SEMANTIC SIMILARITY USING WORDNET

K.vinotha [1] M.A. Maria Parimala M.C.A, M.Phil2, Dr.S.Saravanan[3]
[1]M.Phil scholar PG and Research Department of computer science St.Joseph's college of arts and science (Autonomous) cuddalore
[2]Assistance professor PG and Research Department of computer science St.Joseph's college of arts and science (Autonomous) cuddalore
[3]Assistant Professor, Department of Computer Science &Engineering, Annamalai University

Abstract: A web service is a software system, It automatically clustering Web Service Description Language (WSDL) files into web. It is used to reduce the search engine for service discovery. In proposed, uses two semantic approaches to cluster semantic similar services. First one Latent Semantic Analysis(LSA) it is also known as retrieval technique applied to the collection of WSDL files and the second semantic approach is based on WordNet to cluster the similar web serviceThe users as to retrieve relevant web services. As a result, the comparison of WordNet accuracy is better than the Latent Semantic Analysis (LSA).

Keywords: WSDL, Latent Semantic Analysis, Semantic Similarity, WordNet.

I INTRODUCTION:

Machine Learning is the branch of artificial Intelligences, is to get the computer to learn task such as discriminating between objects, similar data to dissimilar one from learning experience. These also called as supervised learning, unsupervised learning, and Reinforcement learning[1] Supervised Learning: Learning is supervised by the training data. If there are large number of supervised learning method they are nearest neighbor classifiers, decision tree, rule – based [1][2]. Unsupervised Learning: Learning is unsupervised learning method is learned by without the training data. Clustering and topic modeling are commonly used unsupervised Learning Algorithms [1][2]. Whereas clustering and classification are the two operations performed in text document. In proposed they where two approaches are used Latent Semantic Analysis (LSA) and WordNet. As WordNet shows the better performance. In Semantic representation has the following characteristic:

1) More Semantics in the representation.
2) Reduce dimensionality

1.1 Web Services:

A web service is a software system that support Machine – to – Machine interaction is in the format of WSDL files. Web services are prescribed by using SOAP, HTTP, and XML.[3] WSDL document look like this:

```
            <definitions>
            <types>
definition of types........
            </types>
            <message>
definition of a message....
            </message>
<portType>
<operation>
            definition of a operation.......
            </operation>
            </portType>
            <binding>
            definition of a binding....
            </binding>
            <service>
            definition of a service....
            </service>
            </definitions>
```

1.2 WordNet:

WordNet is a lexical database for English language and it is created by cognitive science Laboratory of Princeton University under Psychology professor George A. Miller. It grouping a set of words synonyms called synsets. It provides general definition, records the different semantic relation between synonym sets. WordNet is the difference between the nouns, verbs, adjectives and adverbs to follow grammatical rules.[10] The logical structure of WordNet:

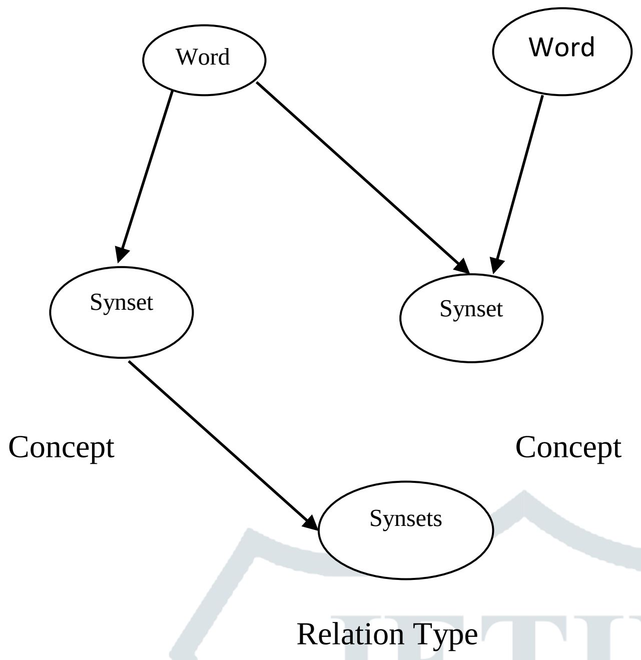


Figure 1.1: The Logical structure of WordNet

II Table of Review

S.NO	Topic	Algorithms	Datasets	Description
1	Classification of Machine Learning Engines using Semantic Indexing	Neural Network K-Nearest Neighborhood and C4.5 algorithm	SourceForge.com Koders.com	The LSI is to classify Neutral network and K-nearest Neighborhood source code. To find the extracting term in source code.C4.5 algorithm give better classifier compare to decision tree.
2	A clustering techniques of news articles using WordNet	K-mean clustering W-K-mean clustering	10,000 news articles from 20 major news like bbc.com,cnn.com etc.	The WordNet used to get a problem like synonymy, ambiguity and lack of content making clusters. The similarity measures in the news articles in web. It

				proposes the two work in wordNet: bag of words in the clustering and label generation. Compare to the result in k-mean.
3	Latent Semantic Indexing: A Probabilistic analysis	-	-	This paper have been compared with document frequency(DF), term contribution(TC), term variance quality(TVQ) this are called unsupervised learning .It improve the accuracy of document clustering.
4	A modification of Wu and Palmer Semantic Similarity measure	WU and PALMER Similarity measure	Physical sensor in the environment, intelligent devices, virtual sensors, Internet access.	This method is to improve the web service discovery process using Jaccard coefficient to calculate the similarity between web services.
5	Survey of clustering Algorithm	k-mean clustering COP-K-mean and Hierarchical clustering	Benchmark dataset-Mushroom and Benchmark dataset-IRIS	This paper is to survey the different clustering algorithm dataset in statistics, computer science and machine learning in some of Benchmark dataset is used to cluster the data for research works.
6	Comparison of Algorithms for document clustering	K-mean clustering and support vector model		This field of data mining, information retrieval have been uses the document clustering, and to compare the different clustering approach, so

				the result is hierarchical based clustering is more efficient than partitioning clustering.
7	A vector space model for automatic indexing	Measure used Precision and Recall	424 document are collected	The VSM is used in the document retrieval or pattern matching (document)is to compared with search request and the property of the customer. These approach are space density based on vocabulary for a collection of documents.
8	Clustering of web services based on semantic similarity	Hierarchical clustering and LERS-M algorithm	Training dataset: Phone number verification services, weather by city	Semantic of web services using WSDL file operation, parameter, wordnet. Is used to similarity in the web service and use data cluster.The test result is used in neighbor approach it give an accuracy level 70%

III PROPOSED WORK:

The Vector Space Model (VSM) is one of the least complex techniques depends on the accurate coordinating of terms that can be found in archives. The cosine of the point between two vectors is utilized as a rule. The outcome is an estimation of comparability running from 0 to 1, where 1 demonstrates an accurate/high match among terms and 0 shows that there is no match. This implies the higher the estimation of the cosine, the higher the probability that two terms are equivalent.[7]

However, the exact matching of phrases raises problems, which includes synonymy and Polysemy. Synonymy deals with special words having the same that means. For instance, automobile and car are synonyms. Polysemy refers to words having more than one distinct that means. which include ―the and ―is, and correlating high similarity measures, result in a excessive match, which does no longer represent the real desired end result

OBJECTIVES:

To overcome the limitations of VSM representation of documents, the proposed work uses two semantic approaches using LSA and Wordnet for clustering similar groups in order to retrieve relevant services for the user query. The main objectives of the proposed work are as follows

1. Use more semantics in the representation
2. Reduce the dimensionality of the feature vectors formed.
3. Improve the cluster quality.
4. Retrieve relevant services to satisfy the user request

The proposed work contains the following three modules.

- Module 1:Keyword-based discovery
- Module 2:LSA- based discovery
- Module 3:Wordnet-based approach

MODULE1: KEYWORD-BASED DISCOVERY:

Keyword-based methods are extensively used in traditional data retrieval structures. An information requester submits the gadget with a question that consists of some of key phrases to be able to retrieve the favored documents.[10] The retrieval system returns stored files in answer to the records requester based at the similarity among the query and the saved files. Here similarity approaches that the files incorporate particular keywords from the requester's query or those files prove comparable enough to the corresponding the question and people files are returned to the records requester.

MODULE2: LATENT SEMANTIC ANALYSIS (LSA):

Semantic similarity approach uses a LSA is similar to the keyword based technique and service description of semantic extraction. The LSA objective has to handle the poor scalability in web and issues in lacking semantics. To achieve these goals , a big service collection is a set f clusters by using k mean clustering algorithm.After the cluster query, the SVD technique is applied to the cluster .[9]

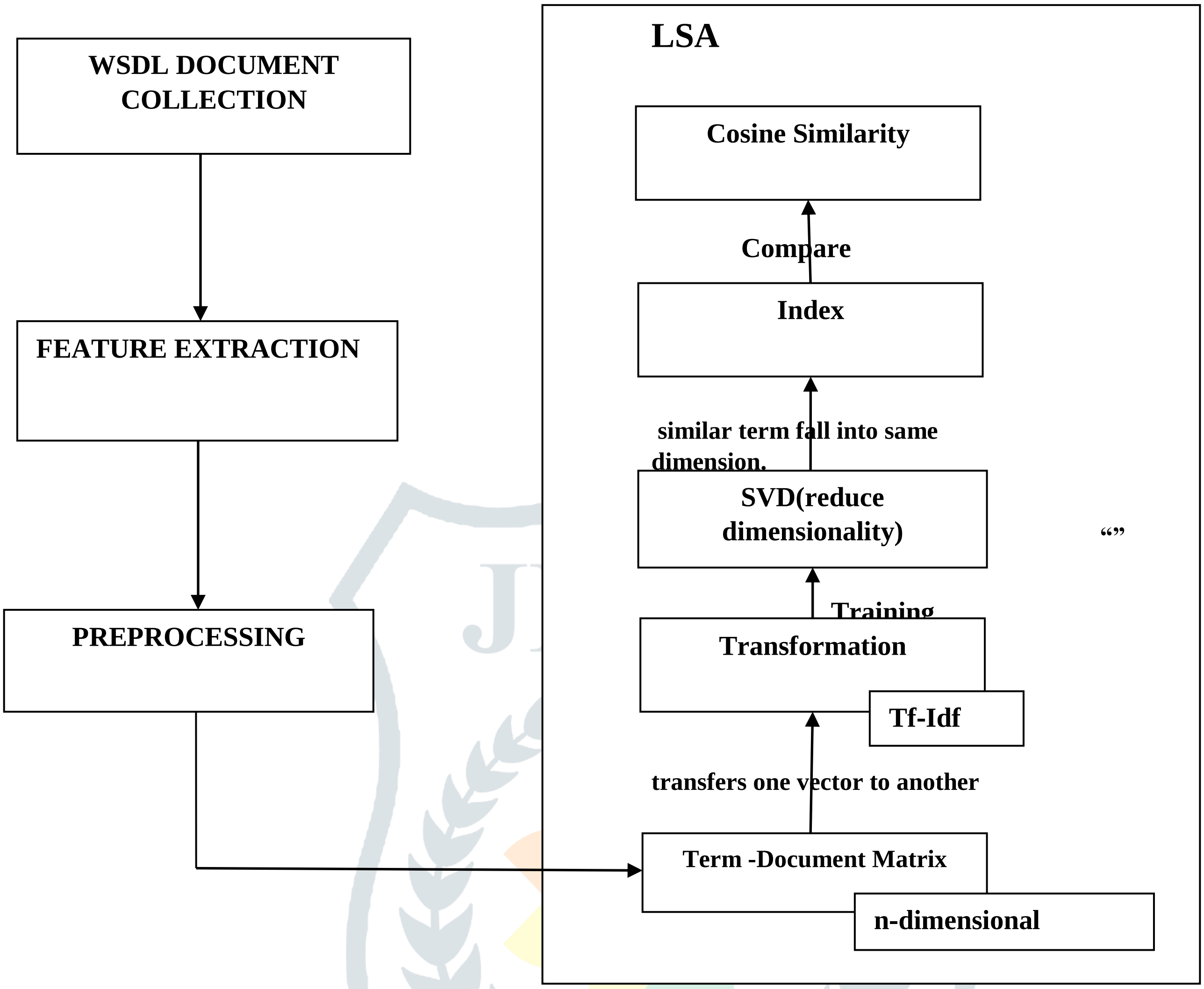


Figure 1.2 The Logical Structure of LSA

SVD (Singular Value Decomposition):

Is a linear algebraic technique that deals with the transformations and the decomposition of the matrix. In that orthogonal transformations uses SVD, matrix is decomposes the products of three sub matrices in information contained in original matrix. This techniques is mostly used in producing a low rank approximation matrix to the initial matrix, for semantics in Information Retrieval(IR). The word document matrix represented as: $A=U\Sigma V^{T}$

MODULE3: SAMANTIC DISCOVERY USING WORDNET

WordNet is uses to find the similarity between words. WordNet can be divided into two methods, they are path length and information content method. Path length method is calculating number of node in taxonomy. Advantage of path length is independent to corpus statical and word distribution. Some of the path length measures are Leacock-Chodorow, Resnik, and Wu-Palmer. In proposed, Wu-palmer equation is implemented for evaluation.

Similarity between two words using WordNet:

Six measure [11]to obtain similarity between words using WordNet. The following measure is based on information content of the least common subsumer (LCS) concepts.

- Leacock-Chodorow
- Resnik
- Wu-Palmer

Three similarity measures is based on path lengths pair of concepts. The proposed work use Wu and Palmer similarity measure [12].Wu and Palmer's method[13] is to calculates similarity to depths of two concepts in WordNet hierarch, with the lowest super-ordinate:

$$sim_{WP}(c_1,c_2)=\frac{2\times depth(lso(c_1,c_2))}{len(c_1,lso(c_1,c_2))+len(c_2,lso(c_1,c_2)+2\times depth(lso(c_1,c_2))} \qquad (1)$$

$$=\frac{2\times depth(lso(c_1,c_2))}{depth(c_1)+depth(c_2)}$$

Similarity based on web services WordNet:

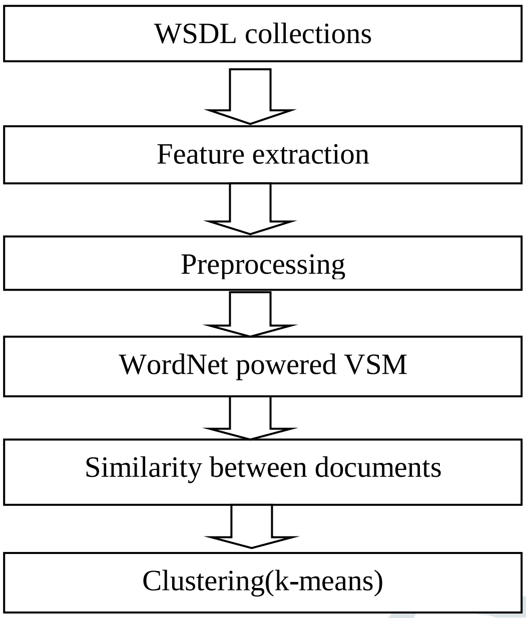
The web service or WSDL files are characterized by name, description and set of operation into input parameter and returns output parameter. In proposed uses the following WSDL information for compute similarity of web services.

Operation name: The web services name is describes by operations performed by a web service.

Message name: The message name describes the input and output parameters for the operation.

Service name: The service name describes the web service.

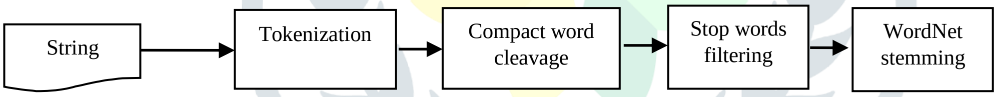
The Architecture describes the work flow of the proposed system. The WordNet based clustering



The collection of 228 real world WSDL files are collected publicly available web services by crawling web services information from www.SOA.Trader.com.

Text preprocessing:

First the service name,operation name, message name, port type name, and port name .The text preprocessing is the first stage it helps to remove unwanted words in filtering process. The following are the steps in process



Procedure of the preprocessing .

Tokenization: In English words, this step is used to remove the white space and non-alphanumeric character.

Compact word Cleavage: The services, is named in Pascal or Camel form of important information of services. Its need to be cleaved. For example:”RealTimeMarketData” it split into “real time marker data”.

Stop word filtering: The repeated words are filtered in this step like “a”,”the”,”an” it should improve the precision.

Stemming: It is used to in removing term suffixes and reducing all forms in stemmed form is useful for recall.

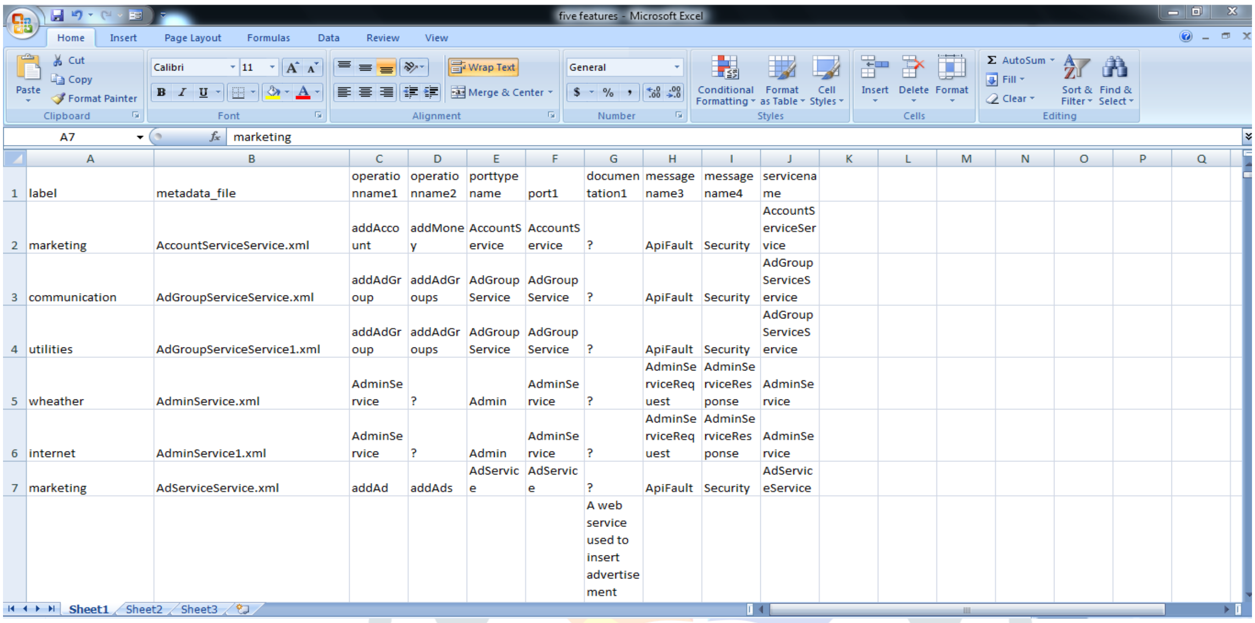
Similarity measures used in WordNet: The similarities between two strings are calculated based on the cosine similarity measure. The formula [14]for cosine similarity,

$$SIM_{cosine}(X,Y) = \frac{A.B}{\|A\| \quad \|B\|}$$

where X, Y are the two service names; A, B are the tokens of the service names X and Y respectively.

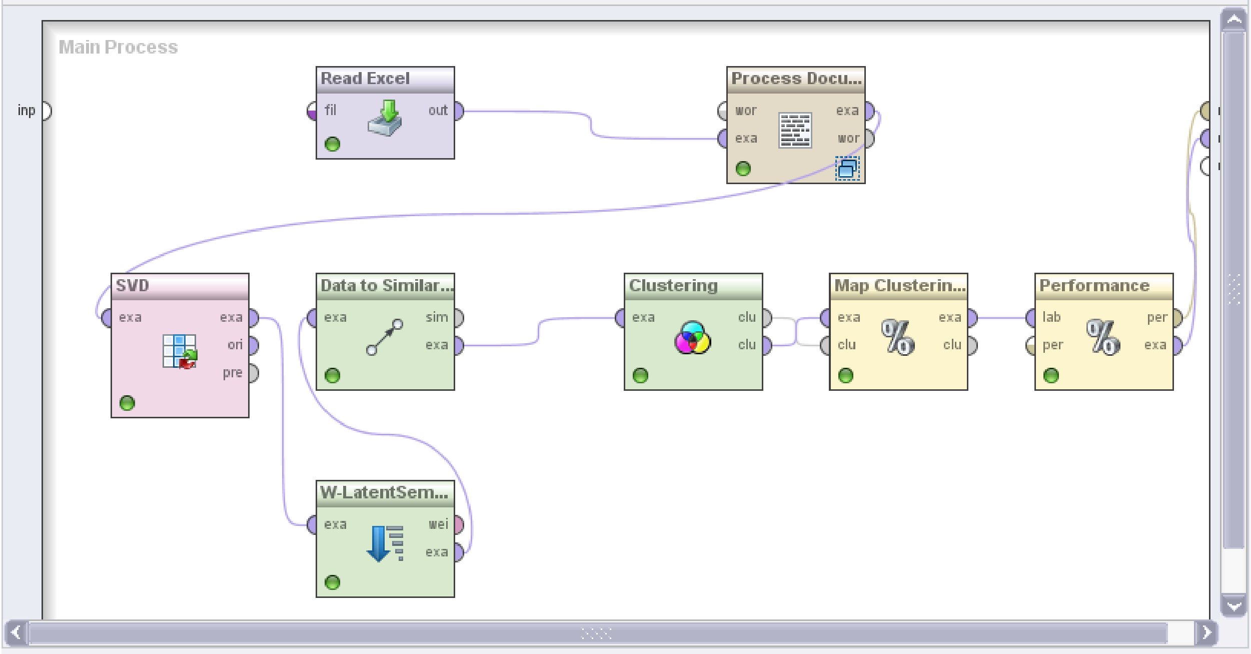
IV EXPERIMENTAL VERIFICATION

Dataset analysis: In excel format:

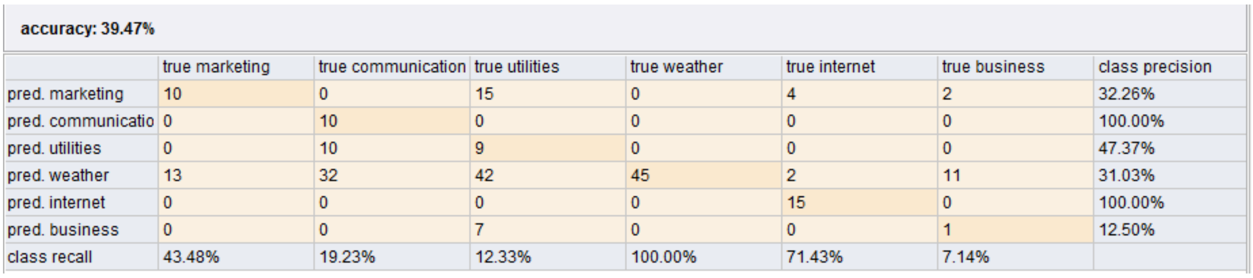


	A	B	C	D	E	F	G	H	I	J
1	label	metadata_file	operationname1	operationname2	porttypename	port1	documentation1	messagename3	messagename4	servicename
2	marketing	AccountServiceService.xml	addAccount	addMoney	AccountService	AccountService	?	ApiFault	Security	AccountServiceService
3	communication	AdGroupServiceService.xml	addAdGroup	addAdGroups	AdGroupService	AdGroupService	?	ApiFault	Security	AdGroupServiceService
4	utilities	AdGroupServiceService1.xml	addAdGroup	addAdGroups	AdGroupService	AdGroupService	?	ApiFault	Security	AdGroupServiceService
5	wheather	AdminService.xml	AdminService	?	Admin	AdminService	?	AdminServiceRequest	AdminServiceResponse	AdminService
6	internet	AdminService1.xml	AdminService	?	Admin	AdminService	?	AdminServiceRequest	AdminServiceResponse	AdminService
7	marketing	AdServiceService.xml	addAd	addAds	AdService	AdService	?	ApiFault	Security	AdServiceService
							A web service used to insert advertisement			

Latent semantic Analysis(LSA)



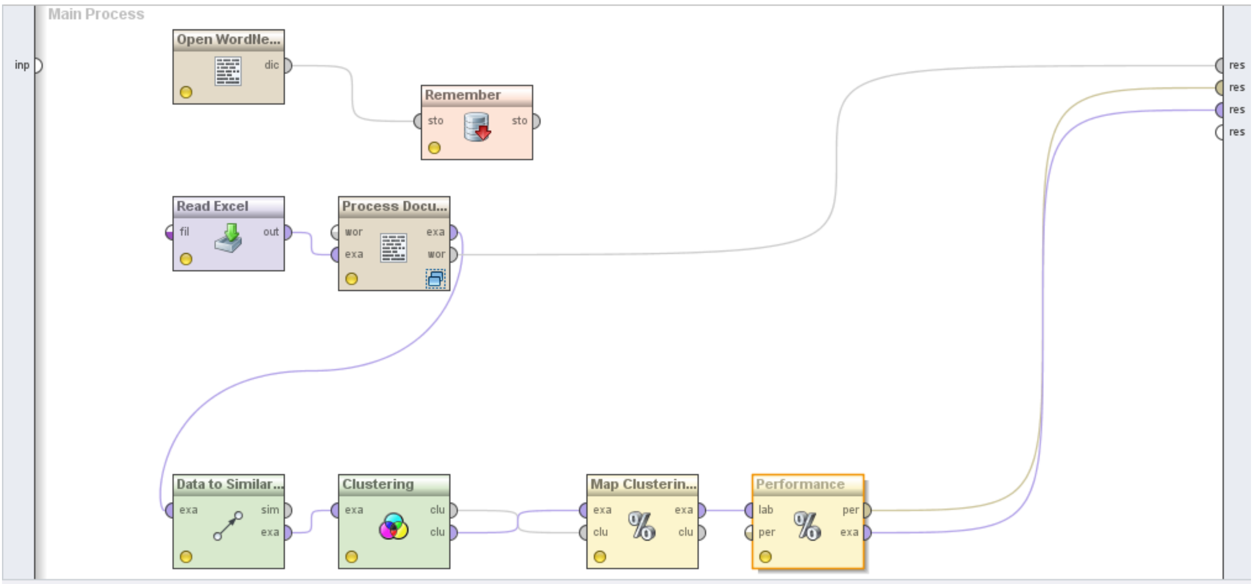
Accuracy of LSA:



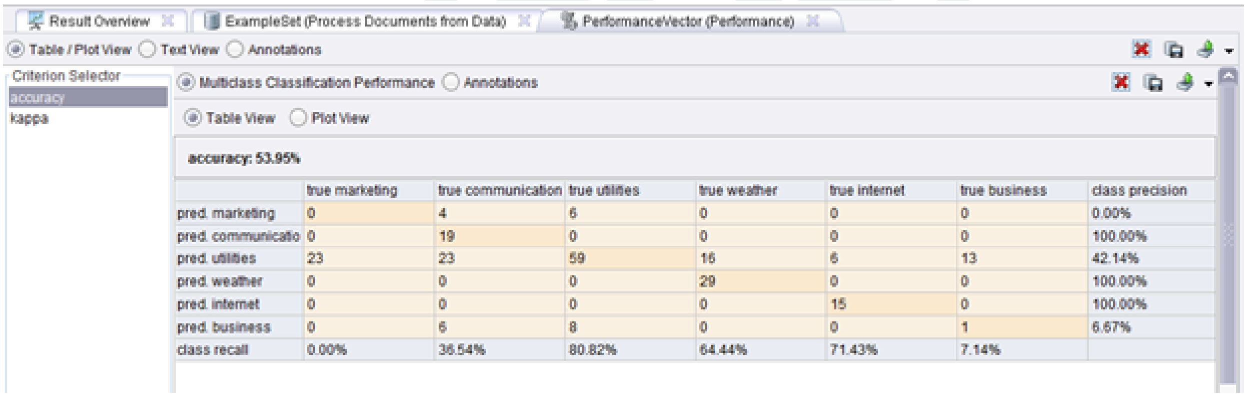
accuracy: 39.47%

	true marketing	true communication	true utilities	true weather	true internet	true business	class precision
pred. marketing	10	0	15	0	4	2	32.26%
pred. communicatio	0	10	0	0	0	0	100.00%
pred. utilities	0	10	9	0	0	0	47.37%
pred. weather	13	32	42	45	2	11	31.03%
pred. internet	0	0	0	0	15	0	100.00%
pred. business	0	0	7	0	0	1	12.50%
class recall	43.48%	19.23%	12.33%	100.00%	71.43%	7.14%	

WordNet:



Accuracy of WordNet:


Result Overview
ExampleSet (Process Documents from Data)
PerformanceVector (Performance)
● Table / Plot View ○ Text View ○ Annotations
Criterion Selector
accuracy
kappa
● Multiclass Classification Performance ○ Annotations
● Table View ○ Plot View

accuracy: 53.95%

	true marketing	true communication	true utilities	true weather	true internet	true business	class precision
pred. marketing	0	4	6	0	0	0	0.00%
pred. communicatio	0	19	0	0	0	0	100.00%
pred. utilities	23	23	59	16	6	13	42.14%
pred. weather	0	0	0	29	0	0	100.00%
pred. internet	0	0	0	0	15	0	100.00%
pred. business	0	6	8	0	0	1	6.67%
class recall	0.00%	36.54%	80.82%	64.44%	71.43%	7.14%	

V RESULT AND DISCUSSION

Proposed work uses two evaluation criteria performance, they are Precision and Recall. Precision can been display the accurate term and Recall is completeness of the term[31]and Precision and Recall values can be used in information retrieval system[32]In proposed is based on comparison of two semantic approaches :

- **Latent semantic Analysis**
- **WordNet**

Result:

1. Comparison between keyword based and wordNet
2. Comparison between keyword based ,LSA and WordNet

Comparison between keyword based and wordNet:

1 Clustering Quality:

The table shows the Average F-measure value for Keyword-based and wordnet to cluster the services. Both of them compared , while WordNet gives the better results,

Approaches	**F-measure**
Keyword-based	32.42
WordNet	36.28

Clustering quality

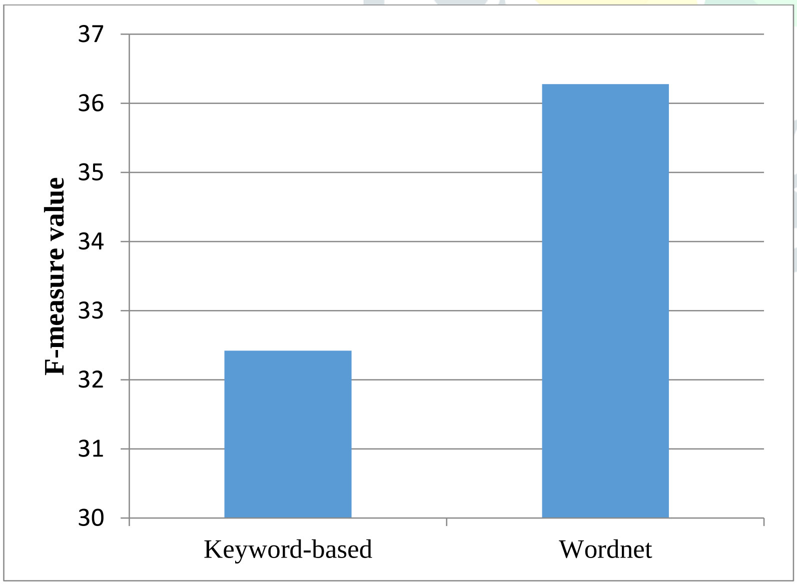


Fig: Comparison of Keyword-based and wordnet

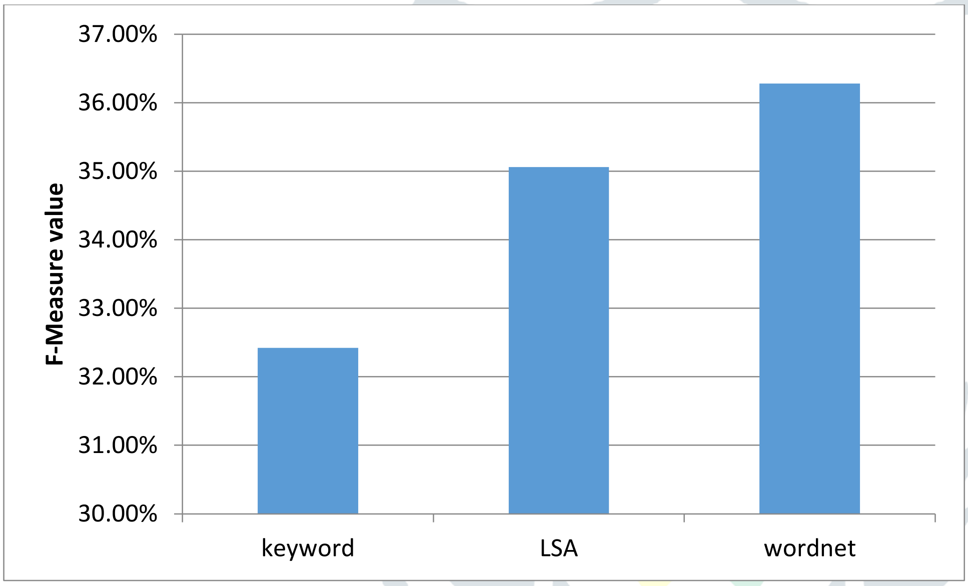
Comparison between keyword based, LSA and WordNet:

Clustering Quality

Comparison between keyword based, LSA and WordNet: The comparison between three approaches for clustering purpose has been implemented, the WordNet gives the better results

Approaches	**F-measure**
Keyword-based	32.42
LSA	35.06
WordNet	36.28

Quality of clustering



Quality of Clustering graph

VI CONCLUSION:

Clustering web services is used to group similar text to reduce the search space .three different approaches are used, keyword based approach it uses the similarity measure, it does not hidden concepts. To reduce the drawback of this techniques,LSA shows better performance compared to keyword based approach.LSA approach uses decomposition techniques, but it does not shows semantic relation between concept. For clustering, WordNet give better performance compare to keyword based approach and LSA.

VII FUTURE WORK

In Future it can be extended to using PLSA(Probabilistic Latent Semantic Analysis) Which as more semantic similarity in document representation.

VIII REFERENCE

[1] Mehdi Allahyari, Seyedamin Pouriyeh and Mehdi Assefi, “A Brief survey of Text Mining:Classification, Clustering and Extraction Techniques”,no.02919, 2017.

[2] Ms.K.Mouthami,Ms.K.Nirmala Devi and Dr.V.Murali Bhaskaran “Sentiment Analysis and classification Based on Textual Reviews”

[3]Aparna Konduri “Clustering of Web Services based on semantic similarity”,2018.

[4] P. Liang, “Opinion Mining on Social Media Data,” pp. 91–96, 2013.

[5] V. Dhanalakshmi and D. Bino, “supervised learning algorithms,” pp. 1–5, 2016

[6] C. Networks and M. Gupta, “COMPARISION OF ALGORITHMS FOR CLUSTERING,” pp. 542–546, 2014.

[7] G. Salton, A. Wong, and C. S. Yang, “A Vector Space Model for Automatic Indexing,” vol. 18, no. 11, 1975

[8] R. Xu, S. Member, and D. W. Ii, “Survey of Clustering Algorithms,” vol. 16, no. 3, pp. 645–678, 2005.

[9] [1]Y. Yusof, T. Alhersh, M. Mahmuddin, and A. M. Din, “Classification of Machine Learning Engines using Latent Semantic Indexing,” no. July, pp. 4–6, 2012.

[10] C. Bouras and V. Tsogkas, “Knowledge-Based Systems A clustering technique for news articles using WordNet ,” *KNOWLEDGE-BASED Syst.*, 2012.

[11]. Pedersen, Ted, Siddharth Patwardhan & Jason Michelizzi (2004), “WordNet::Similarity -- Measuring the relatedness of concepts”, *Demonstrations of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics*, Boston, Mass., 2--7 May 2004, pp. 267--270. Available online at http://www.cs.utah.edu/~sidd/papers/PedersenPM04b.pdf

[12]. Simpson, Troy and Dao, Thanh (2005), WordNet-based semantic similarity measurement, Retrieved 5/25/07 from

http://www.codeproject.com/cs/library/semanticsimilarityWordNet.asp

[13].Wu & Palmer similarity measure, Retrieved 6/1/07 from http://search.cpan.org/src/SID/WordNet-Similarity-

1.04/lib/WordNet/Similarity/wup.pm