Global Journal of researches in engineering
Mechanical and mechanics engineering
Volume 12 Issue 2 Version 1.0 March 2012
Type: Double Blind Peer Reviewed International Research Journal
Publisher: Global Journals Inc. (USA)
Online ISSN: 2249-4596 Print ISSN:0975-5861

MedhAMshaShOdhini – An Efficient Bilingual Search Engine Tool for Telugu Language

By Dr.K.V.N.Sunitha
G.Narayanamma Institute of Technology and Science, India

Abstract - Search engine technology has become quite popular to help users seek information available on the web. The success of a searching system is determined by the quality and efficiency of the search results. There may be very good items on the search topic in other language, but, search engine will generally retrieve items of only one language. Most of these search engines use pattern search which is not efficient. In this paper we present a tool that addresses this problem. Here we discuss the work carried out in developing an efficient tool that retrieves all the items of the database relevant to search term, not just the term matching. This tool retrieves all the synonym matches from both languages. 'MedhAMshashOdhini', meaning, the one which searches exactly what your brain wants to search for, retrieves the documents in both Telugu and English languages. This tool gives the flexibility of searching based on the context, based on the semantics in two different languages. Main claim of the paper is the efficient architecture of thesaurus and the procedure used in retrieving the relevant documents in both languages based on the context.

Keywords : MedhAMshashOdhini, Telugu dictionary, Foreign key, EngTelMap, context resolution, reverse mapping.

GJRE-A Classification : FOR Code: 080704

Strictly as per the compliance and regulations of:

MedhAMshaShOdhini – An Efficient Bilingual Search Engine Tool for Telugu Language

Dr.K.V.N.Sunitha

Abstract - Search engine technology has become quite popular to help users seek information available on the web. The success of a searching system is determined by the quality and efficiency of the search results. There may be very good items on the search topic in other language, but, search engine will generally retrieve items of only one language. Most of these search engines use pattern search which is not efficient. In this paper we present a tool that addresses this problem. Here we discuss the work carried out in developing an efficient tool that retrieves all the items of the database relevant to search term, not just the term matching. This tool retrieves all the synonym matches from both languages. 'MedhAMshashOdhini', meaning, the one which searches exactly what your brain wants to search for, retrieves the documents in both Telugu and English languages. This tool gives the flexibility of searching based on the context, based on the semantics in two different languages. Main claim of the paper is the efficient architecture of thesaurus and the procedure used in retrieving the relevant documents in both languages based on the context.

Keywords : MedhAMshashOdhini, Telugu dictionary, Foreign key, EngTelMap, context resolution, reverse mapping.

I. Introduction

The growth of the Web leads to high popularity of the online search services. The success of a searching system is determined by the quality and efficiency of the search results. Most users have been trained to become accustomed to the traditional search interface where a user submits a query to an input textbox. It is not easy for Web users to specify their search intentions by term combination. Sometimes a user can not generate a query correctly even though the user is clear about what to search.. The vastness of knowledge available on the WWW is the cause of our ever-increasing vulnerability to "not the best" knowledge available. As the number of indexable pages on the Web exceeds, it becomes more and more difficult for search engines to keep an up-to-date and comprehensive search index, resulting in low precision and low recall rates. Users often find it difficult to search for useful and high-quality information on the Web using general-purpose search engines, especially when searching for information on a specific topic or in a language other than English[1]. Many domain specific or language specific search engines have been built to facilitate more efficient searching in different areas. But for Indian languages, it is still not much advanced. For Telugu language, perhaps, there is no efficient search engine that is based on thesaurus.

Telugu has a vast and rich culture and literature dating back to many centuries[2]. Yet there is no widely available electronic thesaurus till date. However bilingual dictionaries are available. For NLP applications, Thesaurus is to be constructed from these dictionaries. The existing searching system gives results only for the search term given and it doesn't provide any specific meanings for the given word depending upon the context.

In this paper, we present our work in designing and implementing a software tool that addresses this problem. We will focus on the architectural design of the tool. The rest of the paper is organized as follows. Section 2 reviews related work in search engine development. Section 3 discusses our research objective. In Section 4, we present our proposed system, called "medhAMshashOdhini," that we have developed to help users create specialized search engines in different domains and languages. Section 5 presents sample results.

II. State of the Art

There are many free software tools that provide all of the components of a search engine. Although these toolkits provide integrated environments for users to build their own domain-specific search engines, most of these tools only work for English documents and are not able to process non-English documents, especially for non-alphabetical languages. Only a few of them, such as GreenStone, support multilingual collection building. As a result, most of these tools cannot be used to create digital library for non-English collections. Another problem is that many of these tools do not provide enough technical details, and their components and building steps are tightly coupled. As a result, users often find it difficult to customize the tools or reuse the intermediate results in other applications (such as document classification) even if they have strong technical skills. For example, in Alkaline, all the intermediate results of the spidering and indexing processes are hidden from users. In addition, most of these tools store the spidering results (Web pages) and indexing results (indexed terms and the document–term relationships) in binary format or other proprietary formats, often due to performance issues. This has made it very difficult, if not impossible, for users to use

Author : CSE Dept, G.Narayanamma Institute of Technology and Science, Shaikpet, Hyderabad, India.
E-mail : k.v.n.sunitha@gmail.com

the results from a search tool for other purposes or in other applications.

The popular search engine, Google, uses the random walk algorithm, which ranks the documents according to the link structure, coupled with the local query specific score to give the final rank to the page[4].

Search engines like oingo.com, excite.com and simpli.com also provide meaning based searching. Launched in October 1999, Oingo has already introduced three fully functional products: DirectSearch, DomainSense and AdSense. DirectSearch, a meaning-based search technology, uses the company's ontology to provide more precise and effective search results. DomainSense, Oingo's meaning-based domain name suggestion technology, currently increases domain name sales for leading registrars around the world. AdSense serves the most highly targeted advertisements on the Internet; effectively targeting advertisements based on search meanings rather than keywords[3].

Step in the direction of meaning based searching was taken by a project of Chinese Academy of Sciences. This project [4], accomplished in 1998, worked on simplified Chinese and English. This project provided the flexibility of adding Traditional Chinese (Big5) and Traditional Chinese (EUC) in the future. Its established system consisted of two subsystems: Organization based subsystem and Web page based subsystem. Organization based subsystem was developed specially for users to find whether a certain organization in China had its own website and some detailed information about that organization. The web page based subsystem was developed for users for searching information in the web documents.

III. MedhAMshaShOdhini

'MedhAMshashOdhini', meaning, the one which searches exactly what your brain wants to search for, retrieves the documents in both Telugu and English languages. This tool gives the flexibility of searching based on the context, based on the semantics in two different languages. Different ways of searching provided by medhAMshashOdhini is described as below.

a) Synonym based Searching

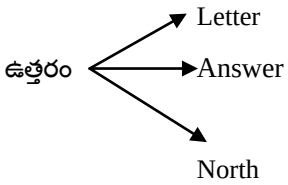
There is no controlled vocabulary or list of standardized terms or descriptors for the Web. Because there is no controlled vocabulary, the use of synonyms and variations of keywords to describe the search query is very important.

b) Context based Searching

If the word has different meanings, depending upon the context all the different meanings will be displayed.

c) Cross linear Searching

For a given term in Telugu we can retrieve words in English. Searching is possible for both the languages i.e searching is possible from English–Telugu as well as from Telugu–English.



The objectives of this work are :

- Manual Construction of Telugu Thesaurus
- Efficient creation and manipulation of the database for the thesaurus.
- Creating transliterable input control so that user can give input in both Telugu and English.
- Using database for context resolution
- Fetching the results using the URL from web server
- Performing web content mining for retrieving the links
- Displaying list of Websites that matches the given search criteria

IV. Architecture of MedhAMshaShOdhini

a) Thesaurus Building

In the proposed system we use relational database(tables) to store and retrieve large amounts of data. Here we consider two schemas, one for generic context and another for specific context. Each context will in turn have two tables. General context contains two tables: GenWordsEng(table for English words), GenWordsTel(table for Telugu words). Specific context contains two tables: EngSp(Spefic context table for English words), TelSp(Specific context table for Telugu words).

i. *GenWordsEng Table*

This table contains all the English words. The table 'GenWordsEng', shown in Table 1, has two attributes: 'Word' and 'Id'. 'Id' is the primary key, which uniquely identifies the words in the table. 'Word' attribute is also unique, which doesn't allow having duplicate copies of word in the table. For each word in the table a unique id is assigned so that, using the id, words can be retrieved from other table. Both will act as the candidate keys. 'Id' of GenWordsTel acts as foreign key for GenWordsEng table. This foreign key is used in EngTelMap as discussed in section 4.4

Word	Id
Letter	1
Rent	2
Hundred	3
Tax	4
Gold	5
Hand	6
answer	7
do	8
reply	9
……	….

Table 1 : English words general context – GenWordsEng

ii. *GenWordsTel Table*

This table contains all Telugu words and its corresponding English word Id. Just like GenWordsEng table, it also has two attributes 'Word' and 'Id' with a difference that in this table, only the attribute 'Word' is unique and hence is the primary key. Id stores the value of GenWordsEng Id. So it may be duplicated. For example, consider word 'Gold'. It's Id in GenWordsEng is 5. All its semantically equivalent words of Telugu will have the same value, i.e. 5, in their Id field as shown in Table 2.

Word	Id
ఉత్తరం	1
లేఖ	1
జవాబు	7
స్వర్ణం	5
బంగారం	5
పైడి	5
పసిడి	5
కనకం	5
పుత్తడి	5
…	…

Table 2 : General Context Telugu table - GenWordsTel

b) Mapping for General Context

Mapping is the main module that retrieves semantically equivalent words from Telugu if the search term is in English and vice versa. As stated above, id of GenWordsEng table is the foreign key of GenWordsTel table. For example, if the search term is 'Gold' whose id is '5' in GenWordsEng, the tool searches for value '5' in Id column of GenWordsTel and retrieves all the rows of id '5' to get the respective synonyms for English word in GenWordsEng. Thus, by this mapping we can get multiple synonyms for given English word as shown in Fig.2

Gold = రం (sva-rNM), బంగారం (baM-gA-rM), పైడి (pai-Di), పసిడ్డి స్వ్ణ (pa-si-di), కనకం (ka-na-kM), పుత్తడి (pu-tta-Di)

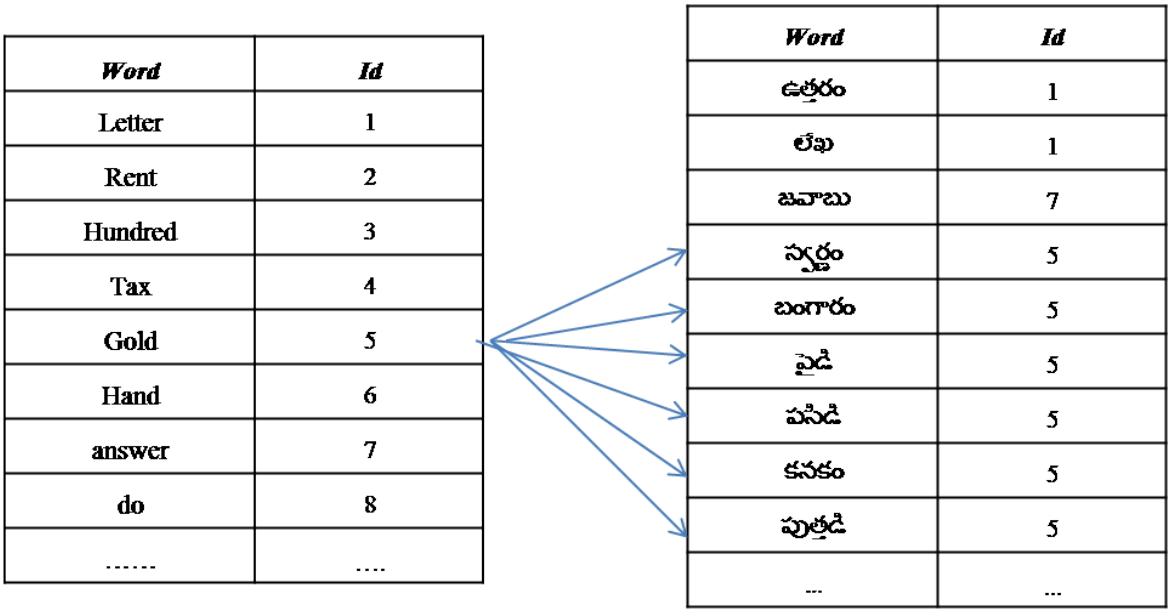
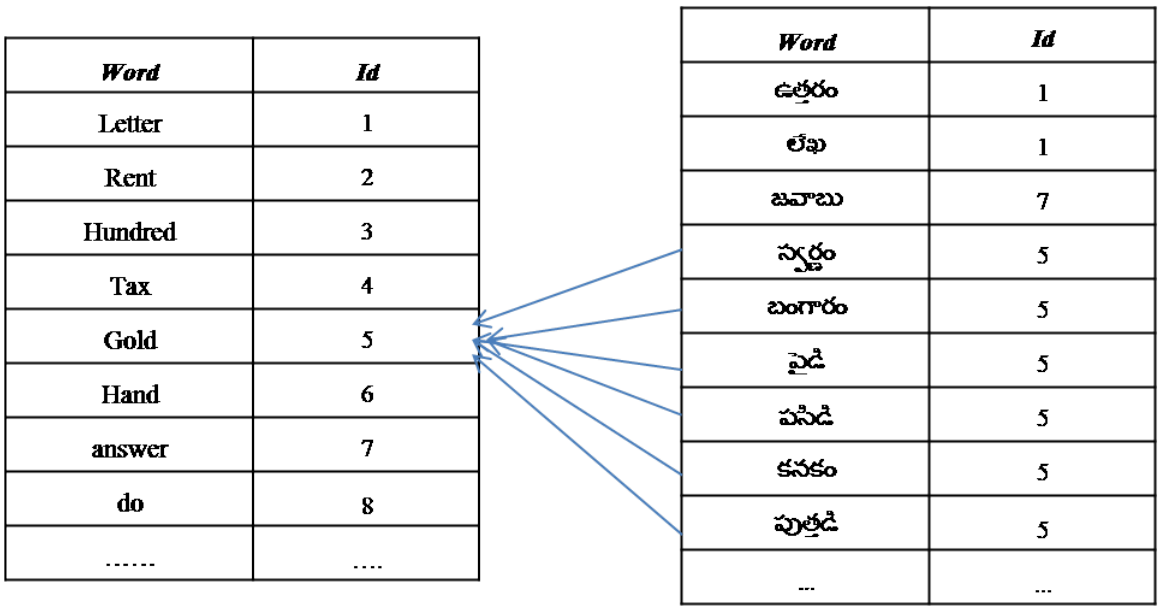


Fig 2 : Mapping word with its synonyms



Word	Id
Letter	1
Rent	2
Hundred	3
Tax	4
Gold	5
Hand	6
answer	7
do	8
.....	....

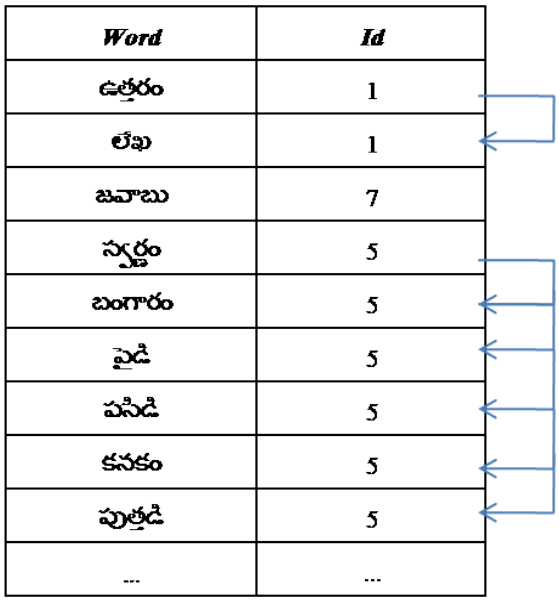
Word	Id
ఉత్తరం	1
లేఖ	1
జవాబు	7
స్వర్ణం	5
బంగారం	5
పైడి	5
పసిడి	5
కనకం	5
పుత్తడి	5
...	...

Fig 3 : Reverse mapping

Fig 3 shows reverse mapping, in which we can get English synonym for given Telugu word. So, if you give any of the above mentioned Telugu word, it will map to the English word 'Gold' with the help of Id field.

స్వర్ణం, బంగారం , పైడి , పసిడి, కనకం, పుత్తడి = Gold

In fig.4, we can observe, mapping within the table, which gives multiple Telugu synonyms for given Telugu word. Given a word in Telugu, get the corresponding Id of it and then scan the whole table for that Id in Id column and extract all the words having that Id. Then the result set contains all the synonyms for the given word.



Word	Id
ఉత్తరం	1
లేఖ	1
జవాబు	7
స్వర్ణం	5
బంగారం	5
పైడి	5
పసిడి	5
కనకం	5
పుత్తడి	5
...	...

Fig 4 : Mapping in the same table

For example, consider the word 'స్వర్ణం', Id of it is'5', searching the whole table for Id 5, we get the words like: బంగారం , పైడి , పసిడి , కనకం, పుత్తడి . Thus in this kind of mapping we can get English/Telugu synonyms for given words of same language.

c) Specific Context Tables and Mapping

These tables are designed for user specific match. A word may be having many meanings. Depending on the user choice, it will search only for those user selected meanings instead of all semantically related words.

i. *EngSpTable*

As shown in Table 3, EngSp table contains four attributes: Id, Eword, Eid, Tid. 'Id' is the primary key which uniquely identifies the word in the table. 'Eword' is the English word which is the unique key avoiding duplicate copies. 'Eid' and 'Tid' are the ids of the English and Telugu synonyms of the word in the same table and the other table. 'Eid's are the Id's of English words in the same table. 'Tid's are the Telugu word id's in table 'TelSp'. Using these two columns id's we get synonyms for the given word.

Id	Word	Eld	Tld
1	reply	7	1,3
2	add	3,4	10
3	join	2,4	10
4	combine	2,3	10
5	Gold	---	5

6	reason	8	11
7	answer	1	1,3
8	excuse	6	11
9	north	---	1
10	grow	---	12
…	…	…	…

Table 3 : Specific context English table (EngSp)

ii. *TelSp*

Table 4 is 'TelSp' table which has Telugu words with its corresponding Ids. This table contains four attributes: 'Id', 'Tword', 'Eid', 'Tid'. 'Id' is the primary key which uniquely identifies the Telugu word in the table. 'Tword' is the Telugu word which is unique, avoiding duplicate entries of the same word. 'Eid' and 'Tid' are the Ids of the English and Telugu synonyms of the word in the same table and other table. 'Eid's are the ids of English words in the EngSp table. 'Tid's are the Telugu word ids in the same table. Using these two columns we get synonyms for the given word.

Id	*Word*	*TId*	*EId*
1	ఉత్తరం	2,3	1,7,9
2	లేఖ	1	1
3	జవాబు	1	7
4	స్వర్ణం	5,6,7,8,9	5
5	బంగారం	4,6,7,8,9	5
6	పైడి	4,5,7,8,9	5
7	పసిడి	4,5,6,8,9	5
8	కనకం	4,5,6,7,9	5
9	పుత్తడి	4,5,6,7,8	5
10	కూడు	---	2,3,4
11	సాకు	12,13	6,8
12	పెంచు	11	10
13	కారణం	11	6

Table 4 : Specific context Telugu table (TelSp)

d) EngTelMap - Mapping

This mapping is done when the given term is in English, and has many meanings; we need to search for a specific meaning. Fig.5 shows the EngTel Mapping. From the figure we can observe the mapping between the two tables in specific context. Given an English word get the Tids of the word. Now scan the 'Id' column of 'TelSp' using the Tid's we got previously, as shown in the figure. So, for example, for the English word 'reply', Tids are:1,3. So, get the words from TelSp table with ids 1 and 3. The resulting set of words is list of synonyms of the word 'reply'.

reply: ఉత్తరం, జవాబు. This illustrates forward mapping from English to Telugu.

Id	*Word*	*EId*	*TId*
1	reply	7	1,3
2	add	3,4	10
3	join	2,4	10
4	combine	2,3	10
5	Gold	—	5
6	reason	8	11
7	answer	1	1,3
8	excuse	6	11
9	north	—	1
….	……	….	

Id	*Word*	*TId*	*EId*
1	ఉత్తరం	2,3	1,7,9
2	లేఖ	1	1
3	జవాబు	1	7
4	స్వర్ణం	5,6,7,8,9	5
5	బంగారం	4,6,7,8,9	5
6	పైడి	4,5,7,8,9	5
7	పసిడి	4,5,6,8,9	5
8	కనకం	4,5,6,7,9	5
9	పుత్తడి	4,5,6,7,8	5
10	కూడు	—	2,3,4
11	సాకు	—	6,8

Fig 5 : EngTelMap

Id	Word	EId	TId
1	reply	7	1,3
2	add	3,4	10
3	join	2,4	10
4	combine	2,3	10
5	Gold	—	5
6	reason	8	11
7	answer	1	1,3
8	excuse	6	11
9	north	—	1
....	.....	....	

Id	Word	TId	EId
1	ఉత్తరం	2,3	1,7,9
2	లేఖ	1	1
3	జవాబు	1	7
4	స్వర్గం	5,6,7,8,9	5
5	బంగారం	4,6,7,8,9	5
6	పైడి	4,5,7,8,9	5
7	పసిడి	4,5,6,8,9	5
8	కనకం	4,5,6,7,9	5
9	పుత్తడి	4,5,6,7,8	5
10	కూడు	—	2,3,4
11	సాకు	—	6,8

Fig 6 : TelEngMap

The figure 6 shows how the backward mapping goes from Telugu table to English table in specific context. Given a Telugu word, first get the Eid of the corresponding Tword from 'Eid' column. Scan the EngSp table for the retrieved ids previously, and extract the respective English words from EngSp table, then the resulting set contains all the English synonyms for the given Telugu word. In Fig 7, we can see the mapping within the table, where we can get the multiple Telugu synonyms in various context for given Telugu word. Given a Telugu word, we get the Tid of the corresponding word and scan the same table for the retrieved id. The result set contains the set of Telugu words which resembles meanings for given word in various contexts.

For eg:word: సాకు = పెంచు, కారణం

Id	Word	TId	EId
1	ఉత్తరం	2,3	1,7,9
2	లేఖ	1	1
3	జవాబు	1	7
4	స్వర్గం	5,6,7,8,9	5
5	బంగారం	4,6,7,8,9	5
6	పైడి	4,5,7,8,9	5
7	పసిడి	4,5,6,8,9	5
8	కనకం	4,5,6,7,9	5
9	పుత్తడి	4,5,6,7,8	5
10	కూడు	—	2,3,4
11	సాకు	12,13	6,8
12	పెంచు	11	10
13	కారణం	11	6

Fig 7 : Mapping in the same table

IX. Conclusion

The overall quality of web searching system is determined not only by the prowess of its searching algorithm, but also by the caliber of its corpus, both in terms of comprehensiveness (e.g. coverage of topics,

language, etc.) and refinement (e.g. freshness, avoidance of redundancy, etc.).The relative corpus size estimates competitive marketing advantage and bragging rights in the context of the web searching.

We have collected 700 Telugu words with their respective synonyms in general context and 25 Telugu words with different synonyms and semantically related words in specific context. In all the corpus has 1,020 words stored in a very efficient manner.

Manual procedures of thesaurus building can be a bottleneck of our proposed approach. In our future work we are going to address the problem by making use of automated lexical acquisition. The automatically constructed thesaurus can also be taken as a starting point for developing a better searching system. In the present system we have developed tools for searching Telugu and English words. This can be extended to other languages so that the user can perform multilingual searching at one go.

References Références Referencias

1. Micheal Chau,et.al.,"SpidersRUs: Creating specialized search engines in multiple languages", Science Direct, Decision Support systems, 45 (2008), pg: 621-640
2. G.Uma Maheshwara Rao, "Morphological complexity of Telugu", ICOSAL-2, 2000.
3. Pushpak Bhattacharya,et.al., " A multi Lingual Meaning Based Search Engine"
4. S.Brin,L.Page, "The anatomy of a Large-Scale Hypertextual Web Search Engine", Proceedings of the7th WWW Conference, Brisbane, Australia, Apr 1998.
5. H.Chen, M.Chau,D.Zeng, CI Spider: a tool for competitive intelligence on the web, Decision Support Systems 34(1) (2002) 1-17
6. J.Lovins. "Development of stemming algorithm", Journal of mechanical translation and computational linguistics, 11:22-31,1968
7. K.V.N.Sunitha, A.Sharada, "Building an Efficient Language Model based on Morphology for Telugu ASR", KSE-1, March 2010, CIIL, Mysore
8. K.V.N.Sunitha, A.Sharada, "Telugu Text Corpora Analysis for Creating Speech Database", IJEIT,ISSN 0975-5292, Dec 2009, Volume 1, No.2
9. Paice C and Husk G. "Another Stemmer". In ACM SIGIR Forum 24(3):566,1990
10. M.F.Porter. "An algorithm for suffix stripping".In readings in information retrieval, pages 313-316, San Francisco,CA,USA,1997.Morgan Kaufmann Publishers Inc.
11. Jinxi Xu and W. Bruce Croft."Corpus based stemming using co-occurrence of word variants". ACM Trans.Inf.Syst., 16(1):61-81,1998
12. J.L.Dawson. "Suffix removal for word conflation". In Bulletin of the Association for Literary and Linguistic Computing, volume 2(3), pages 33-46, Michaelmas,1974
13. R.Krovetz. "Viewing morphology as an inference process". In Proceedings of Sixteenth Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pages 191-203, 1993
14. vishwabharat@tdil
15. Emerald Group Publishing Ltd.,"Issues in Indian languages computing in particular reference to search and retrieval in Telugu Language"
16. "LANGUAGE IN INDIA Strength for Today and Bright Hope for Tomorrow", Vol 6, August 2006