Accessing Unstructured Data through Mobile Devices

I.Vijayalakshmi
Department of Electronics,
Madras Institute of Technology,
Anna University, India

Sobha Lalitha Devi
AUKBC Research Centre
Madras Institute of Technology,
Anna University, India.

ABSTRACT

The paper presents an on-going work on accessing unstructured data in the web through mobile devices. To achieve this we use Information Extraction (IE) to extract relevant information from unstructured documents. Here the relevant information are extracted are stored into a database, where a user can search information by giving a query through mobile. The extracted information that matches with the given query in the database are retrieved and presented in a mobile environment. The information extracted is used for searching through the mobile device. This work is done for two languages, English and Tamil. The search facility is provided for text documents that exists in the mobile as well as on the web and the results are presented in the mobile.

General Terms

English grammer, Tamil grammar, Natural Language Processing, Information Extraction.

Keywords

Morphological Analyzer, Tokenizer, Parts of Speech Tagger and chunker, Named Entity Recognizer.

1. INTRODUCTION

Mobiles phones have intruded in our lives and have become an important component for our future strategies. The search provided on the mobile is a developing branch of Information Retrieval is centered on the convergence of mobile platforms, mobile phones and other mobile devices. Web search for mobile phone allows the users to find the contents on websites.

Different types of mobile search are – mobile optimized search engines, mobile directory search engines, mobile discovery engines, and mobile question and answer services. Mobile optimized search engines have implemented an optimized version of bandwidth and form factor in the mobile platform. Some of the mobile optimized search engines are Yahoo and Google. Mobile directory search engines allow the users to find local services in the locality of their current location. Mobile discovery engines services suggest user's recommendations on what they should do next. Mobile question and answer services provide the user to type a question to a central database and the reply is got via text.

To achieve the search facility we need to extract information from the documents by Information Extraction (IE) using Natural Language Processing (NLP). In this several unstructured documents hold valuable data that is used to extract relevant information from the database. The main purpose of using NLP is to design and build software that will understand, analyze and generate languages that humans use naturally, so that the computer will address as though addressing a human being.

2. NATURAL LANGUAGE PROCESSING (NLP) TASKS

The NLP task requires two main engines such as (i) Language Processing Engine, and (ii) Information Extraction Engine. The Language Processing Engine consists of the following steps (a) Parts of Speech (POS) tagging – It marks each word as noun, verb, preposition, adjective, determiner etc., (b) Chunker – It identify the phrases such as NP, VP, PP, ADVP etc., (c) Key term Extractor – These are words which are carrying the action or information (d) Named Entity Recognition (NER) – Named Entities refer to definite noun phrases which have specific types of individuals, such as persons, location, organizations, date, and so on.

The objective of NER system is to identify all the named entities which are in textual mention. The two sub tasks of NER are: (i) identifying its type, and (ii) identifying the boundaries of the NE. Consider an example we want find the answer for query of all companies that are merged recently with the information of their CEOs. In this it uses a join optimization of Information Extraction output [1]. From this we get the named entities such as organization, person name. The IE Engine consists of various rules that is followed for different types of documents that is used for extraction from the database .The following topics cover the rules used for extraction from the TE Engine.

3. SYSTEM ARCHITECTURE

The Search Engine (SE) module has been divided into several components. Each component is being assigned to do a specific task involved in the SE. The components are

- *Preprocessing:* The preprocessing contains the trained data sets which take the plain text as input which is moved to the preprocessor and get the trained data sets which are given as input to the IE Engine.
- *IE Engine:* IE Engine is governs the trained data set to identify the domain specific key word and extract relevant information.
- *IE Output:* The extracted information from the IE Engine moved to the database and the output is stored in that database.
- *IR Engine:* IR (Information Retrieval) Engine is used for retrieve appropriate data from the IE Output.
- When a user enters a query into the mobile, it is used to move to the IR Engine search information from the IE database and display the results on the output window.

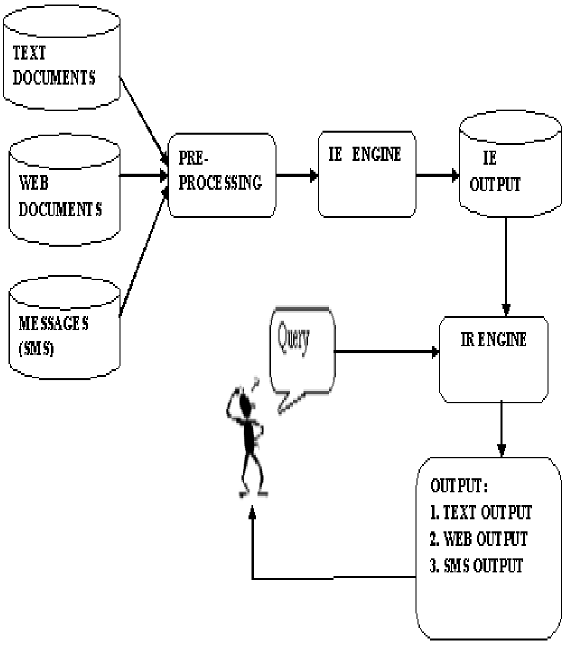


Fig 1. Block diagram for search engine from mobile phones

3.1 IE ARCHITECTURE

The plain text is given as Input to the IE Engine for preprocessing and gets Information Extraction from the preprocessed text. First, the plain text is split into sentences using a sentence segmenter, which further subdivided into words using tokenization. Next, each tokenized sentences is tagged with POS tagging, it marks the noun, verb, preposition, determiner according to the sentences given with respect to the input. Next, the POS tagged sentences moved to the Chucker where it identifies the phrases like NP, VP, and PP etc... Next, the NER used to tell what type of Named Entity is used such as Person, organizations, date, time and so on[4]. The main purpose of using NER is to identify the named entities that exist in the textual format. Finally, the Relationship Extraction is used to find the most likely relation that exists in the text document and extract the output related to the relation where the relation between each text in the entities are different. The IE architecture is given below:

3.2 MORPH ANALYZER

The basic analysis of any Natural Language Processing task is Morphological Analysis. The study of internal structure of the word is called Morphology. The noun and verb morph analyzer for different tasks such as noun is used for handling nouns and proper nouns, the verb is used to handle all verb forms like auxiliary, finite and infinite forms. It use Romanized word for segmenting the Tamil graphemes. The implementation of morphological analysis consists of preprocessing, morpheme Segmentation, and morph syntactic tagging [2]. The morphological analyzer consists of preprocessing, morphological segmentor is done by tokenization, POS tagging, Chunking, NER is found and the Information Extraction is done for Tamil documents [3].

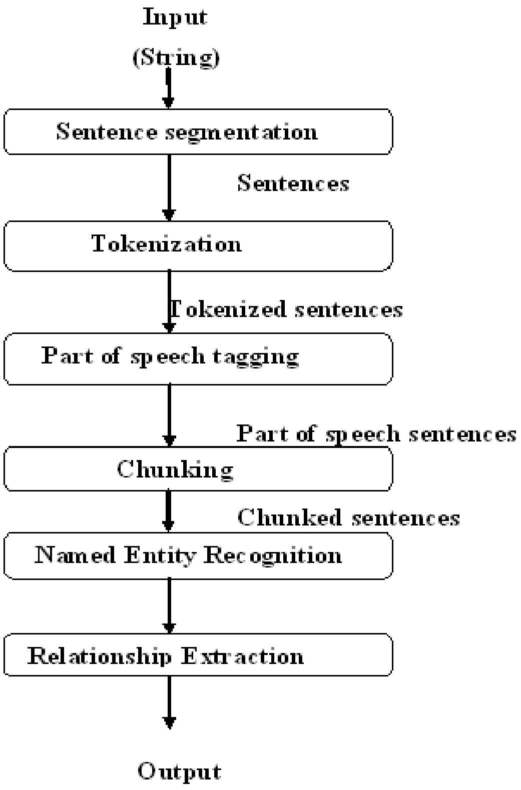


Fig 2. IE Architecture

4. ALGORITHM FOR EXTRACTED RELATION

The Information Extraction algorithm is one of the key factors affecting the result quality. The extraction will attempt to meet the user-specified quality requirements as efficiently as possible. All the domain specific keywords and rules corresponding to each domain specific keywords are initialized first. Next, the preprocessed file is read for each sentence. Next, the domain specific keywords are collected and stored. Next, the rules are checked for sentences that satisfies the rules in the database. If the rules are matched it extract the fields and exhibit the result, if it doesn't match with the rule it is used to check for other sentences that existing from the data base.

```
Query:  agaya gangai
Output:  waterfall details
      - Input is moved to the pattern matching
      -  The regular expression matches the given pattern

       Pattern p = Pattern. Compile ("Rules for IE Engine");
     - check whether the regular expression matches to the input
        Sentence
          Matcher m = p.matcher (input);
   -Using while command, find the keywords from the pattern
       Matched sentences

                  While (m.find ())
   -After the output is got from the pattern matcher and stored the
keywords are matched for the extraction needed and the Information
Extraction output is got from the .txt files
```

Fig 3. Algorithm for Information Extraction

The algorithm shown in figure 3 is used for extracting relevant information from unstructured documents from the database. After preprocessing, the preprocessed input is taken for extraction where it matches with the IE rules for against the pattern matching. If the pattern matches with the given regular expression it is moved to the while loop condition and there information extraction is done by extracting related information from the text and show the result.

4.1 INFORMATION EXTRACTION RULES

Information Extraction rules helps us to extract the information for different data that exists in the database. Some of the rules are as follows.

Rule 1:

{NEs_location, facilities} [sym] VP_OPEN [V01] 001001 VP_CLOSE [B01] P01 NP P03 {NE_distance} {P09 NE_location}

where,

NEs_location	- The Named Entity such as place
[Sym]	-Symbols such as [**:/:],[)]**, etc
VP_OPEN	-It used for opening of the verb phrase.
[V01]	- It gives verb, 3rd person singular Present
001001	- It gives the verb past tense.
VP_CLOSE	- It is used for closing the verb phrase.
P01, P03, P09	- It gives the conjunction, subordinating.
{NE_distance}	- It is used to give the distance of the place.

Rule 2:

{NE_location} C04 {NE_location} VP_OPEN V01 002001 VP_CLOSE P04 {NPs}

where,

NEs_location	- The Named Entity such as place
C04	- It gives the coordinating Conjunction.
VP_OPEN	-It used for opening of the verb phrase.
[V01]	- It gives verb, 3rd person singular present
002001	- It gives the verb past tense.
VP_CLOSE	- It is used for closing the verb phrase.
P04	- It gives the conjunction, subordinating.
{NPs}	- It is used to give the noun phrase.

5. RELEVANT INFORMATION EXTRACTION FROM UNSTRUCTURED DOCUMENTS

A user gives a query for getting relevant information from an unstructured document. By using the IE rules we extract the relevant information that satisfies the requirements. Here, the Information Extraction is done for waterfalls in India. Suppose the unstructured document contains the information in English.

5.1 English documents:

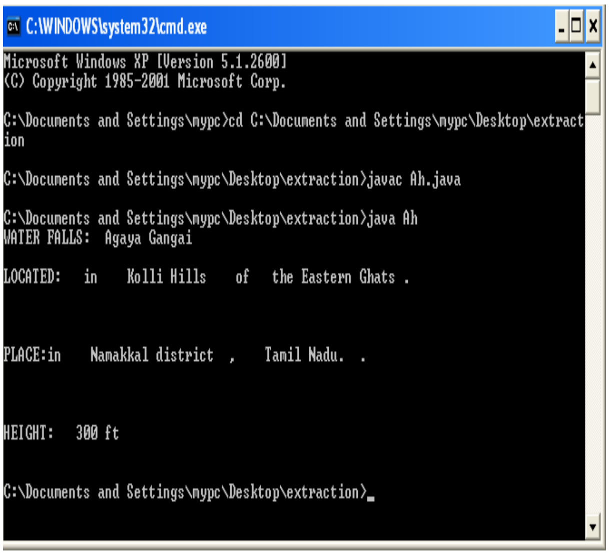
Agaya Gangai waterfalls is located in Kolli Hills of the Eastern Ghats. It is in fact a multi-tiered waterfall and one can see different shades of this waterfall through its various tiers from different view points. Panchanathi, a jungle stream cascades down as the Agaya Gangai (En:Ganges of Sky), near the Arapaleeswarar temple atop the Kolli Hills in Namakkal district, Tamil Nadu.[1] It is 300 ft waterfall of the river Aiyaru situated close to Arapaleeswarar temple. It is located in a valley that is surrounded by mountains on all sides. [2] The entire terrain is green and hence the temperatures are slightly lower than outside. Though, the humidity can get slightly high.

Access

Steep set of thousand plus steps from the Arapaleeswarar temple takes one to the foot of the waterfall. The descent from the temple to the waterfall should take about 15 minutes while the ascent should take about 20-25 minutes. Agaya Gangai falls is an ideal place to take a bath. Close to the waterfall is a small room where one can keep their baggage and enjoy a shower under the waterfall. The rocks close to the waterfall were slippery and hence we had to watch our feet.

Fig 4: Raw plain text before Information Extraction

The results for the text document shown in figure 4 for Agaya gangai after extracting relevant data from the unstructured document is shown in figure 5:

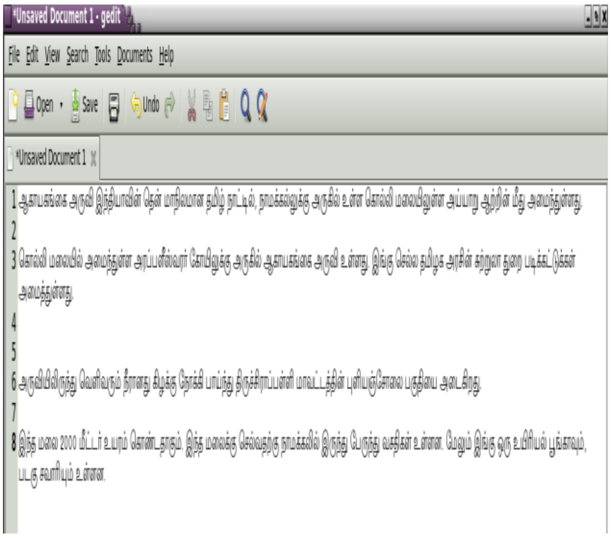


```
Microsoft Windows XP [Version 5.1.2600]
(C) Copyright 1985-2001 Microsoft Corp.

C:\Documents and Settings\mypc>cd C:\Documents and Settings\mypc\Desktop\extract
ion

C:\Documents and Settings\mypc\Desktop\extraction>javac Ah.java

C:\Documents and Settings\mypc\Desktop\extraction>java Ah
WATER FALLS:  Agaya Gangai

LOCATED:   in    Kolli Hills    of   the Eastern Ghats .



PLACE:in    Namakkal district  ,    Tamil Nadu.  .



HEIGHT:   300 ft


C:\Documents and Settings\mypc\Desktop\extraction>_
```

Fig 5. Extracted Plain text for Agaya gangai in English

5.2 Tamil documents:

The plain text in unstructured format for Tamil documents before extracting relevant information is shown in figure 6. For the same query the output in Tamil is shown in figure 7.


*Unsaved Document 1 - gedit
File Edit View Search Tools Documents Help
Open
Save
Undo
*Unsaved Document 1

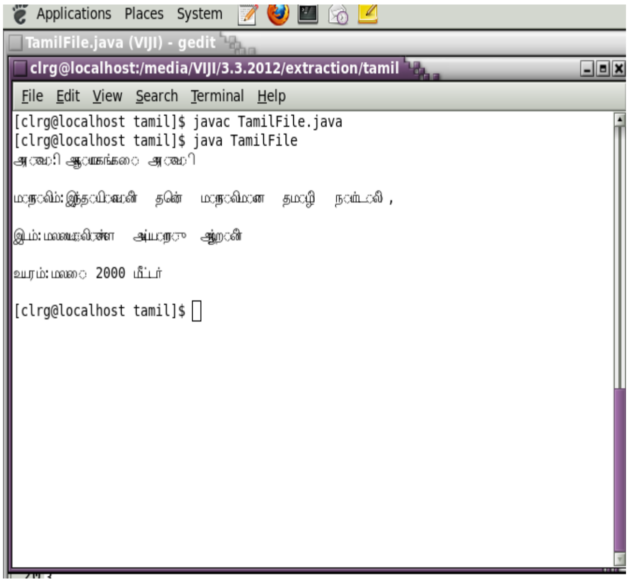
ஆகாயகங்கை அருவி இந்தியாவின் தென் மாநிலமான தமிழ் நாட்டில், நாமக்கல்லுக்கு அருகில் உள்ள கொல்லி மலையிலுள்ள அய்யாறு ஆற்றின் மீது அமைந்துள்ளது.

கொல்லி மலையில் அமைந்துள்ள அரப்பளீஸ்வரர் கோயிலுக்கு அருகில் ஆகாயகங்கை அருவி உள்ளது. இங்கு செல்ல தமிழக அரசின் சுற்றுலா துறை படிக்கட்டுக்கள் அமைத்துள்ளது.

அருவியிலிருந்து வெளிவரும் நீரானது கிழக்கு நோக்கி பாய்ந்து திருச்சிராப்பள்ளி மாவட்டத்தின் புளியஞ்சோலை பகுதியை அடைகிறது.

இந்த மலை 2000 மீட்டர் உயரம் கொண்டதாகும். இந்த மலைக்கு செல்வதற்கு நாமக்கலில் இருந்து பேருந்து வசதிகள் உள்ளன. மேலும் இங்கு ஒரு உயிரியல் பூங்காவும், படகு சவாரியும் உள்ளன.

Fig 6. Plain text before Information Extraction


Applications Places System
TamilFile.java (VIJI) - gedit
clrg@localhost:/media/VIJI/3.3.2012/extraction/tamil
File Edit View Search Terminal Help

```
[clrg@localhost tamil]$ javac TamilFile.java
[clrg@localhost tamil]$ java TamilFile
அருவி: ஆகாயகங்கை அருவி

மாநிலம்: இந்தியாவின் தென் மாநிலமான தமிழ் நாட்டில் ,

இடம்: மலையிலுள்ள அய்யாறு ஆற்றின்

உயரம்: மலை 2000 மீட்டர்

[clrg@localhost tamil]$ 
```

Fig 7. Extracted information for Agaya gangai in Tamil.

The Information Extraction from the text and web documents displays the same result. Since, the web pages stored in the database are first converted to text file. Then preprocessing was done for those web documents and displays the output. Like this, around 200 documents are taken and got the output from the database.

6. CONCLUSION AND FUTURE WORK

The extraction is done for the system level for the user query. The Information Extraction is done for a domain for each key word exists in the current sentence and collects the rule that matches the specific key word. The sentence are checked whether it satisfies the rule, then the words which matches the specific rule are matched and filled to get the desired output through Information Extraction.

The Information Extraction is done for extracting specific key words which from the sentences satisfies the rules have to be extracted and filled in the template. The work has to be improvised for more specific key words in the templates, meet the constrains in the mobile and adapted to the mobile environment supporting for English and Tamil.

7. REFERENCES:

[1] Alpa Jain, Panagiotis G. Ipeirotis, AnHai Doan, Luis Gravano, Join Optimization of Information Extraction Output. Columbia University, New York University, University of Wisconsin-Madison.

[2] Anand Kumar M, Dhanalakshmi V, Soman K.P,2010. A Sequence Labeling Approach to Morphological Analyzer for Tamil Language. Computational Engineering and Networking (CEN), AMRITA Vishwa Vidyapeetham,Coimbatore, India.

[3] Dhanalakshmi V, Anand Kumar M, Soman K.P, CEN, Amrita Vishwa, 2010. Natural Language Processing Tools for Tamil Grammar Learning and Teaching.Vidyapeetham Coimbatore, India.

[4] Fabio Ciravegna. Adaptive Information Extraction from Text

[5] by Rule Induction and Generalisation. Department of Computer Science, University of Sheffield, Regent Court, 211 Portobello Street, S1 4DP Sheffield, UK

[6] S. Lakshmana Pandian, T.V. Geetha, 2009. Semantic Role Labeling for Tamil Documents. Department of Computer Science and Engineering, ANNAUNIVERSITY, Chennai, India.

[7] Raymond J. Mooney and Un Yong Nahm. Text Mining with Information Extraction. Raymond J. Mooney and Un Yong Nahm

[8] Department of Computer Sciences, University of Texas, Austin, TX 78712-1188.

[9] http://gate.ac.uk/