Advances in Computational Research
ISSN: 0975-3273 & E-ISSN: 0975-9085, Volume 4, Issue 1, 2012, pp.-125-129
Available online at http://www.bioinfo.in/contents.php?id=33

ITRANSED MARATHI LITERATURE RETRIEVAL USING SMS BASED NATURAL LANGUAGE QUERY

PATHAK V.M.[1]*AND JOSHI M.R.[2]

[1]Department of MCA, IMR Jalgaon Affiliated to NMU Jalgaon, MS, India
[2]Department of Computer Science North Maharashtra University, Jalgaon, MS, India
*Corresponding Author: Email-

Received: February 21, 2012; Accepted: March 06, 2012

Abstract- SMS based information retrieval is one of the challenging research driven application of information and communication technology. "A flexible natural language query in the form of an SMS could also be answered by applying information retrieval theory and practices", is our research theme. In the initial phase, we have formulated a preparatory experimentation on SMS based Marathi query to retrieve Marathi literature documents.
This paper presents the details about the ITRANS Marathi literature documents followed by the description of an appropriate Vector Space Model (VSM) to represent the documents and the queries. We designed a GUI and asked a group of hundred students to formulate an SMS query for 30 different questions in order to obtain user specific variations among typed in queries. A collection of sample Marathi documents in the ITRANS form and a set of acquired queries is processed using the system developed to retrieve relevant documents. The system uses the Cosine Similarity to rank the documents according to the relevance for each selected query. Experimental results of the system are presented in the paper. This is the first such type information retrieval model for the ITRANS Marathi documents.
Keywords- Information retrieval, ITRANSE code, Vector space model, Cosine Similarity, SMS based IR.

Citation: Pathak V.M. and Joshi M.R. (2012) ITRANSED Marathi literature retrieval using SMS based natural language query. Advances in Computational Research, ISSN: 0975-3273 & E-ISSN: 0975-9085, Volume 4, Issue 1, pp.-125-129.

Introduction

The facilities offered by mobile devices now include the possibility to retrieve information using SMS/IM messages. Major web search engines such as Yahoo and Google are improving their search facility to propose such services for mobile search [7]. "Personalized information access", could be termed as the major characteristic where Mobile based information retrieval differentiates from traditional Information Retrieval (IR) [7]. Personalization refers to the information access as per the user preferences as well as the query formulation context. This needs more precise and refined query expressions to improve the overall retrieval effectiveness of the service. In India the spoken languages and their respective scripting make the major attribution of the "personalization of information" problem. With increasing digitization of libraries, literatures of various local languages from different states, are being restored in digitized form by applying innovative transliterations and encoding formats. The other side of the coin is that, our young generation is practicing over transliterated local language SMS to exchange and share their thoughts on mobiles with their friends. Combining these two sides of the same coin as "information retrieval and SMS services", can be considered as an evolutionary step for an effective ICT implementation in coming age.

SMSbIR, a Literature Survey.

With above background and interest to understand the present scenario on SMS based information retrieval in various application domain, a survey was designed, executed and published by authors of this paper in IEEE conference proceeding indexed on "IEEE-Explore" [6].

This survey has been extended further to explore more dimensions by studying more systems with wide range applications. In this study we found that most of the work is based on planned fixed form SMS query to access information from static data bases or dynamic web search facility using WAP. Some researchers are coming up with new ideas for relevant information retrieval using SMS facility [5]. Still applying SMS based flexible query in natural language to access information from a well structured knowledge base need to be explored in many angles. This could open a new horizon in domains such as education, medical, tourism, governance and to a great extent in library domain.

Motivation.

When most of the mobile users are using SMS facility of their hand sets to exchange information, images, thoughts, jokes and many other text and non-text entities, using this facility for better ICT applications becomes the obvious research problem. Natural language SMS could be send by the users to a specified service using mobile's wireless connectivity, where the service could have access to some knowledge source as per the domain in question to answer the query in the same natural language is the core concept of the SMS based information retrieval [4].

When a survey was done for library domain, the members of a local library irrespective to age and gender have given their 100% support to the concept of having library access on their mobile hand set. Encouraged after these initial surveys, "a remote access to the literature in library in digitized form on mobiles" is selected as the underline domain of SMS based information retrieval to do our research experimentations.

About ITRANS Documents

When explored through the encoding methods applied to restore Marathi literature in digitized form, the Indic-transliteration (ITRANSE) form came across as the best transliteration coding format. This scheme encodes most of the Indic-language documents in self tagged English script form, including character coding for all vowels, consonants, special symbols such as "Anusvara", "Avagraha", "Visargah", "Chandra/ArdhChandra bindu" and much more. The concept is that the documented contents can be retrieved and printed in the local language script form such as "Devanagari", for Marathi language. Any exertion to retrieve such ITRANSED documents using natural language query is not still on record of IR. With the fact that most of the students and young generation in India use English script to type SMS in local languages (like Marathi language in Maharashtra, Gujarati in Gujarat state), the author of this paper started working on experimenting the problems related to "SMSbIR" in this domain. SMS will be formed in Marathi language-English script style using flexible query forms. This paper is about the experiments performed to collect a few pre-formed Marathi queries in various forms from students of computer application and management of author's institute. About 36 students in this experiment have submitted about 260 queries from a sample 30 various predesigned queries hand- written in Marathi. They are asked to type them in English script by applying what so ever the short forms, they use for SMS. Further these queries are refined to remove the noise to prepare corrected queries in our next step of experimentation. This next step of experimentation includes design and application of cosine similarity to score the relevance of documents with the refined queries. Presently the queries are manually refined and used in experiment to identify problems of applying Cosine similarity using, "Vector Space Model" on ITRANSE Marathi literature retrieval. As an example of ITRANSE Marathi literature a poem of "Bahinabai Chaudhari" has been presented in Appendix of this paper.

Scope and Aim

To initiate the development of an SMS based information retrieval system we have first started working on the query collection. The Marathi literature documents of poems, auto biography of great personalities, novels, historical events, current affairs in various fields and much more could be made accessible by interfacing between the SMS service and information retrieval models. Many researchers and developers are coming up with innovative application domains with this concept. This related work has been compiled to build a taxonomy of SMS based IR systems by the authors. It is first necessary to select an effective encoding method of digitized library literature of Marathi. In our study in this direction we found Indic transliteration format called ITRANS. This encoding method restores most of the Indian language literature including Marathi, Malayalam, Guajarati, Bengali and many others with a base of Sanskrit documents, in English script. Our first aim is to extract contents of these ITRANSed Marathi documents (Language Marathi but English script), as per the query send as SMS in Marathi language with English script.

Thus the objective of the study, that we present in this paper are as follows.

- To study possible variations in the query terms for the same query as submitted by different users.
- To study the query refining strategy for the SMS based queries in Marathi with English script.
- To study the applicability of Vector Space Model and Cosine Similarity.

Layout of the paper

After introduction of the problem we have identified in first section, the section two explains the experiments we have carried out to collect raw queries and processing to make it relevant to the document's database. The possible variations in the query terms of the same Marathi query, is presented and query refining is done in this section. Section four describes the implementation of information retrieval using vector space model for Marathi literature. The experiment designed and executed with the application of cosine similarity for relevance ranking of the documents is discussed here. The last section submits the conclusion remarks for the problem that has been worked out.

Table 1- Query processing and refinement.

Submitted query	Corrected query as in ITRANS
mn vdhay2 kavite che kavi kon	Mana vaDhAya vaDhAya kaviteche kavi koN
"Raghunandan savata" chi gayika kon?	Raghunandan sa.vaLa chi gayika koN
"ailama pailama" he bhondlyache gane havi	ailamA pailamA he bho.nDalyAchl gAnil.n have
Odumbr che kv kn	auduMbara che kavi koNa

Query collection and processing.

The problem undertaken for this first part of our experimentation can be defined as-

"To identify types of noise involved from user specific variation in the collection of Marathi query set using SMS facility".

Methodology

We designed and executed an experiment on students of computer application and Management in two different batches respectively. 30 pre-formed Marathi queries were handwritten in "Devnagari" script. Students were provided with 10 queries out of these 30 on papers. The required Software has been developed under guidance of author of this paper by MCA students using "JAVA Socket- Programming", on Windows-XP in the computer lab. 18 students of the course of Computer Application in first experiment submitted 100 queries in given time limit. By giving corrected instructions with the experience of first experiment, second experiment was successfully performed with help of Computer Management students, to receive near about 160 queries from these 18 students in given time limit. This made the total figure of Marathi in English queries, in the database as 260. As the client software demanded for the students' mobile numbers, the queries from different students could be differentiated. We present here the snap shot of this client software.

The server running on one PC collected the queries into a data base from clients, simultaneously receiving the submitted queries from students in the computer laboratory. The students are now encouraged to develop the system using mobile's SMS interfacing with the server for the experiments in future. Here the snap shot of this server's software is presented in Fig 2. It gives us idea of the query collection. The variation in noise in the queries submitted by different students for the same query could be easily pointed out from this snap shot. A query processing is thus a necessary part of the IR implementation.

Noise types.

When the collection of 260 queries is analyzed we have detected the variations in the same query from different participants. The noise involved in these variations could be specified in types such as Shortcuts, Contractions, Misspells, SMS Clippings, Abbreviations, Word duplication, Numerated words.

Query Processing

From the collected 250 such queries, we selected near about 75 queries for further processing. The query processing includes validation, parsing and term matching. In addition to this SMS queries require for removal of noise using SMS Normalization and query refining [1] to make the queries allocable for the purpose of Information Retrieval. Validation checks if any invalid character or term occurs in the text as per the domain of the application. Parsing separates the terms as per the grammar / syntax of the sentences in the respective language. We have used space as the delimiter of the terms. Extra spaces are removed between two consecutive query terms. Then the query terms are corrected from the words of document corpus. Presently this has been managed manually to select refined queries for further experimentation. The query refinement will be part of the IR and relevance feedback model that is the research domain of the author under guidance of the co-author.

From the snap shot of Query submission server we would list few examples of the query submitted and the corrected query after refining it as per ITRANS form, in the table below. From such 75 variations of the pre-formed 30 Marathi queries we have selected 16 queries from the database, refined them and used them for the application "Cosine Similarity". This experiment is named as "COSIM" for further reference.

Fig. 1- Snap Shot of Query Submission Client

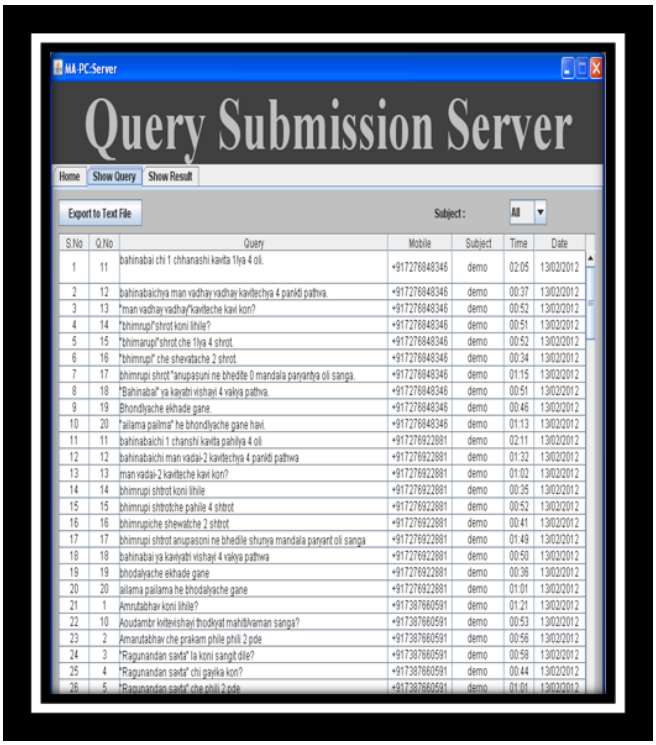

Fig. 2 - Snap Shot of "Query Collection Server"

VSM and COSIM experiment

In order to initiate our work as defined previously we searched for the "Source of Corpus" for ITRANS Marathi documents. We found number of ITRANS documents freely available as .txt, .pdf and mainly .itx files on the internet [www.sanskritdocuments.org]. Another site is www.aczoom.com; that is authored by "Avinash Chopade". This site gives historical information of ITRANS encoding. An online and downloadable ITRANS interface is also available as free-wares on these sites. We have selected 40 ITRANS Marathi documents from internet. Out of these only 15 documents were preprocessed to remove stop-words, extra spaces, special ITRANS tags and symbols. With the aim to apply Vector Space Model over these documents and then to apply Cosine Similarity formulation this pre-processing is applied. One sample pre-processed document is presented in Appendix for better understanding of the format ITRANS Marathi literature document. The original document is copied from internet [www.sanskritdocument.org].

The COSIM program is developed by author this paper in JAVA that reads the documents, and queries to create vectors as Hash-sets. If the same term appears in the document the frequency count of the term is increased. The next part of program reads the queries one at a time in a loop to count its term frequency weights in respective document and also frequency weights in the query.

The Cosine similarity formula is applied using summation of cross product of these weights as below.

$$\mathrm{Sim(j/i)} = \frac{\sum_{i=1}^{n} WiDij}{\sqrt{Wi^2}\sqrt{Dij^2}}$$

Where Sim(j/i) shows the similarity of jth docum ent with ith query , W_i is weight of ith term query and Dij is ith query-term weight in document j. The program gives a matrix of Q x D , with matrix element being Sim (j/i) where j is 1 to MAX_D and i= 1 to MAX_Q . The weights are assigned as the frequency of terms in query and documents respectively.

COSIM matrix and analysis

The COSIM experiment was planned to be a first step to apply VSM for relevant information retrieval of ITRANSED Marathi literature. We have selected Marathi poems including "Abhanga", "Stotram", "Povada" and even religious contents such as "HaripaTh". The First iteration was done without any preprocessing of the documents. It worked but has given poor results to highlight a few major problems like different stop word, extra spaces, end of line and end of file detections. Then documents were selected to preprocess them to sort out above problems. In second iteration, the refined query set of 10 queries verses 15 documents as above were processed by COSIM program to produce Queries Vs. Document Matrix. The results are then matched with the ranks manually given by the participants to understand the applicability of VSM and Cosine similarity using its frequency model.

The documents database is processed to form a weighted vector table for each document. Queries are also processed to create similar vector table. Then the Similarity matrix is generated to present it to the participants for document ranking. As for any specific query only one document is relevant as per the sample selection, the highest rank document is checked for the relevance. The participants are asked to assign ranks to 3 relevant documents from high to low as 1 to 3. Thus the results are then analyzed with a precision 1.

Rank Analysis

Table 2- Rank analysis

Query #	1'st rank doc #	Relevance Feedback	Rank by participant
1	14	yes	1
6	7	yes	1
7	15	no	2
12	4	yes	1
15	11	no	2
16	15	yes	1

From above table we can understand that out of 6 query cases 4 are answered correctly with highest rank documents. From the feedback of participants and contents of related documents the ranks are assured. The query # 15 has no relevant document in tested database but it has given document #11 as highest rank document. Query# 7 gives no 15 as highest document where document # 1 is the relevant document.

Conclusion

This is the first experiment over Marathi ITRANS document which has given 66% correct results. In future VSM implementation will be refined to improve the precision and recall values. More documents and queries will be experimented. The query collection will be performed by using SMS on mobile connection. And query processing, document preprocessing will be done based on the findings of this experiments and related term matching and noise removal algorithms.

Appendix

```
\engtitle \{.. Mana Vadhaya \..}##
\itxtitle \{.. mana vaDhAya vaDhAya \..}## \endtitles \##
kavi  bahINAbAI
mana vaDhAya vaDhAya ubhyA pIkAtala.n Dhora
kitI hAkalA hAkalA phirI yeta.n pikA.nvara
mana pAkharU pAkharU tyAchI kAya sA.ngU mAta \##?##
AtA vhata.n bhu_lvara gela.n gela.n AbhAyAta
mana laharI laharI tyAle hAtI dhare kona \##?##
u.nDArala.n u.nDArala.n jasa.n vArA vAhAdana
mana jahyarI jahyarI yAcha.n nyAra.n re ta.ntara
Are \##,## ichU \##, ## sApa barA
tyAle utAre ma.ntara \##!##
devA \##,## Asa.n kasa.n mana \##?##
Asa.n kasa.n re ghaDala.n
kuThe jAgepanI tule
asa.n sapana paDala.n \##!##
\##
```

References

[1] AiTi Aw, Min hang, Juan Xiao, Jian Su, *A Phrase-based Statistical Model for SMS Text Normalization*.

[2] Ademola O. Adesina et. al. (2011) *International Journal of Soft*

Computing and Engineering, 1.
[3] Cedrick Fairo et al. (2009) "*A Translated Corpus of 30,000 French SMS*.
[4] Jeunghyun Byun, Seung-Wook Lee, Young-In Song, Hae-Chang Rim (2008) *AAAI Workshop on Enhanced Messaging*.
[5] Agbele K., Adesina A. (2011) *Research Journal of Information technology*.
[6] Manish Joshi, Varsha Pathak (2011) *IEEE explore*.
[7] Marthe Buffiere Frederic Pichon (2005) *Knowledge Based Flexible Query Answering " a thesis submitted as partial fulfillment of degree of International Master in Information Technology , submitted and supervised by University of ESBJERG*.