AN UNSUPERVISED APPROACH TO DEVELOP IR SYSTEM: THE CASE OF URDU

Mohd. Shahid Husain

Integral University, Lucknow

ABSTRACT

Web Search Engines are best gifts to the mankind by Information and Communication Technologies. Without the search engines it would have been almost impossible to make the efficient access of the information available on the web today. They play a very vital role in the accessibility and usability of the internet based information systems. As the internet users are increasing day by day so is the amount of information being available on web increasing. But the access of information is not uniform across all the language communities. Besides English and European languages that constitutes to the 60% of the information available on the web, there is still a wide range of the information available on the internet in different languages too. In the past few years the amount of information available in Indian Languages has also increased. Besides English and few European Languages, there are no tools and techniques available for the efficient retrieval of this information available on the internet. Especially in the case of the Indian Languages the research is still in the preliminary steps. There are no sufficient amount of tools and techniques available for the efficient retrieval of the information for Indian Languages.

As we know that Indian Languages are very resource poor languages in terms of IR test data collection. So my main focus was mainly on developing the data set for URDU IR, training and testing data for Stemmer.

We have developed a language independent system to facilitate efficient retrieval of information available in Urdu language which can be used for other languages as well. The system gives precision of 0.63 and the recall of the system is 0.8. For this Firstly I have developed an Unsupervised Stemmer for URDU Language [1] as it is very important in the Information Retrieval.

Keywords: *Information Retrieval, Vector Space Model, Stemming, Urdu IR*

1. INTRODUCTION

The rapid growth of electronic data has attracted the attention in the research and industry communities for efficient methods for indexing, analysis and retrieval of information from these large number of data repositories having wide range of data for a vast domain of applications.
In this era of Information technology, more and more data is now being made available on online data repositories. Almost every information one need is now available on internet. English and European Languages basically dominated the web since its inception. However, now the web is getting multi-lingual. Especially, during last few years, a wide range of information in Indian regional languages like Hindi, Urdu, Bengali, Oriya, Tamil and Telugu has been made available on web in the form of e-data. But the access to these data repositories is very low because the efficient search engines/retrieval systems supporting these languages are very limited. Hence automatic information processing and retrieval is become an urgent requirement. Moreover, since India is a country having a wide range of regional languages, in the Indian context, the IR approach should be such that it can handle multilingual document collections.

DOI : 10.5121/ijaia.2013.4506

A number of information retrieval systems are available to support English and some other European languages. Work involving development of IR systems for Indian languages is only of recent interests. Development of such systems is constraint by the lack of the availability of linguistic resources and tools in these languages. The reported works in this direction for Indian languages were focused on Hindi, Tamil, Bengali, Marathi and Oriya. But there is no reported work is done for Urdu language.

There is no sufficient amount of resources available to retrieve information effectively available on internet in Indian Languages. So there is a need of some efficient tools and techniques to represent, express, store and retrieve the information available in different languages.
The present work focuses on development of an efficient Information Retrieval system for Urdu Language.

2.INFORMATION RETRIEVAL

Information Retrieval is the sub domain of text mining and natural language processing. This is the science in which the software system retrieves the relevant documents or the information in response to the user query need. The Information Retrieval system match the given user quires with the data corpus available and rank the documents on the basis of the relevance with the user need. Then the IR system returns the top ranked documents containing relevant information to the user query.

IR systems may be monolingual, bi-lingual or multilingual. The main objective of this thesis work is the development of the mono-lingual information retrieval system for Urdu language.
To retrieve the relevant information on the basis of user query

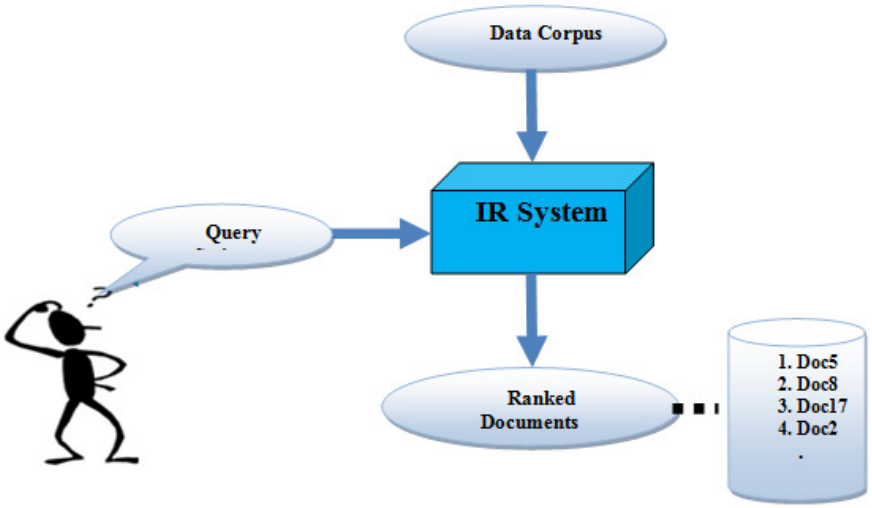

Fig. 1: typical IR system

- The IR system breaks the query statement and the data corpus in a standard format.
- The query is then matched with the documents presented in the corpus and ranked on the basis of the relevance with the query.
- Top ranked documents are then retrieved.

There are various approaches for converting the query statement and the corpus data in a standard format like stemming, morphological analysis, Stop word removal, indexing etc.
Similarly there are various techniques or methods for query matching like cosine similarity, Euclidean distance etc.

The efficiency of any Information Retrieval system depends on the term weighting schemes, strategies used for indexing the documents and the retrieval model used to develop the IR system.

2.1 Stemmer

Stemming is the backbone process of any IR system. Stemmers are used for getting base or root form (i.e. stems) from inflected (or sometimes derived) words. Unlike morphological analyzer, where the root words have some lexical meaning, it's not necessary with the case of a stemmer.
A stemmer is used to remove the inflected part of the words to get their root form. Stemming is used to reduce the overhead of indexing and to improve the performance of an IR system. More specifically, stemmer increases recall of the search engine, whereas Precision decreases. However sometimes precision may increases depending upon the information need of the users. Stemming is the basic process of any query system, because a user who needs some information on plays may also be interested in documents that contain the word play (without the s).

2.2 Term Frequency

This is a local parameter which indicates the frequency or the count of a term within a document. This parameter gives the relevance of a document with a user query term on the basis of how many times that term occurs in that particular document.
Mathematically it can be given as: $tf_{ij}=n_{ij}$
Where nij is the frequency or the number of occurrence of term ti in the document dj.

2.3 Document Frequency

This is a global parameter and attempts to include distribution of term across the documents. This parameter gives the importance of the term across the document corpus. The number of the documents in the corpus containing the considered term t is called the document frequency. To normalize, it is divided by the total number of the documents in the corpus.

Mathematically it can be given as: $df_i=n_i/n$

Where n_i is the number of documents that contains term **ti** and the total number of the documents in the corpus is n.
idf is the inverse of this document frequency.

2.4 The third factor which may affect the weighting function is the length of the document.

Hence the term weighting function can be represented by a triplet ABC

here A- tf component
B- idf component
C- Length normalizing component

The factor Term frequency within a document i.e. A may have following options:

Table 1: different options for considering term frequency

n	tf = tfij	(Raw term frequency)
b	tf = 0 or 1	(binary weight)
a	$tf = 0.5 + 0.5\left(\frac{tf_{ij}}{\max tf \text{ in } D_j}\right)$	(Augmented term frequency)
l	$tf = \ln(tf_{ij}) + 1.0$	Logarithmic term frequency

The options for the factor inverse document frequency i.e. B is:

Table 2: different options for considering inverse document frequency

N	Wt=tf	No conversion i.e. idf is not taken
T	Wt=tf*idf	Idf is taken into account

The options for the factor document length i.e. C is:

Table 3: different options for considering document length

N	W_{ij}=wt	No conversion
C	W_{ij}=wt/ sqrt(sum of (wts squared))	Normalized weight

2.5 Indexing

To represent the documents in the corpus and the user query statement indexing is done. That is the process of transforming document text and given query statement to some representation of it is known as indexing.

There are different index structures which can be used for indexing. The most commonly used data structure by IR system is inverted index.

Indexing techniques concerned with the selection of good document descriptors, such as keywords or terms, to describe information content of the documents.

A good descriptor is one that helps in describing the content of the document and in discriminating the document from other documents in the collection.

The most widely used method is to represent the query and the document as a set of tokens i.e. index terms or keywords.

For indexing a document, there are different indexing strategies as given below :

2.5.1 Character Indexing:

In this scheme the tokens used for representing the documents are the characters present in the document.

2.5.2 Word Indexing:

This approach uses words in the document to represent it.

2.5.3 N-gram indexing:

This method breaks the words into n-grams, these n-grams are used to index the documents.

2.5.4 Compound Word Indexing:

In this method bi-words or tri-words are used for indexing.

2.6 Information retrieval models

An IR model defines the following aspects of retrieval procedure of a search engine:
a. How the documents and user's queries are represented
b. How system retrieves relevant documents according to users' queries &
c. How retrieved documents are ranked.

Any typical IR model comprises of the following:
a. A model for documents
b. A model for queries and
c. Matching function which compares queries to documents.
The IR models can be categorized as:

2.6.1 Classical models of IR:

This is the simplest IR model. It is based on the well recognized and easy to understood knowledge of mathematics.

Classical models are easy to implement and are very efficient.

The three classical information retrieval models are:

-Boolean
-Vector and
-Probabilistic models

2.6.2 Non-Classical models of IR:

Non-classical information retrieval models are based on principles like information logic model, situation theory model and interaction model. They are not based on concepts like similarity, probability, Boolean operations etc. on which classical retrieval models are based on.

2.6.3 Alternative models of IR:

Alternative models are advanced classical IR models. These models make use of specific techniques from other fields like Cluster model, fuzzy model and latent semantic indexing (LSI) models.

2.6.4 Boolean Retrieval model:

This is the simplest retrieval model which retrieves the information on the basis of the query given in Boolean expression. Boolean queries are queries that uses And, OR and Not Boolean operations to join the query terms.

The one drawback of Boolean information retrieval model is that it requires Boolean query instead of free text. The second drawback is that this model cannot rank the documents on the basis of relevance with the user query. It just gives the document if it contains the query word, regardless the term count in the document or the actual importance of that query word in the document.

2.6.5 Vector Space model:

This model represents documents and queries as vectors of features representing terms. Features are assigned some numerical value that is usually some function of frequency of terms.
In this model, each document d is viewed as a vector of tf×idf values, one component for each term
So we have a vector space where

a. Terms are axes
b. documents live in this space

Ranking algorithm compute similarity between document and query vectors to yield a retrieval score to each document. The Postulate is: Documents related to the same information are close together in the vector space.

2.6.6 Probabilistic retrieval model:

In this model, initialy some set of documents is retrieved by using vectorial model or boolean model. The user inspects these documents looking for the relevant ones and gives his feed back. IR system uses this feedback information to refine the search criteria.

This process is repeated, untill user gets the desired information in response to his needs.

2.7.Similarity Measures

To retrieve the most relevant documents with the user information need, the IR system matches the documents available in the corpus with the given user query. To perform this process different similarity measures are used. For example Euclidean distance, cosine similarity.

2.7.1 Cosine Similarity

We regard the query as short document. The documents present in the corpus and the query are represented by the vectors in the vector space with features as axes.

The IR system rank the documents by the closeness of document vectors to the query vectors. IR system then retrieve the top ranked documents to the user.

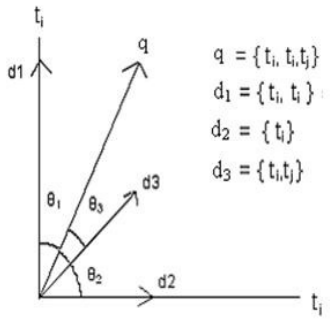

Fig. 2: A VSM model representing 3 documents and a query

The above diagram shows a vector space model where axes t_i and t_j are the terms used for indexing.

The cosine similarity between the document d_j and the query vector q_k is given as:

$$sim(d_j, q_k) = \frac{(d_j, q_k)}{\|d_j\| \|q_k\|} = \frac{\sum_{i=1}^{m} W_{ij} \times W_{ik}}{\sqrt{\sum_{i=1}^{m} W_{ik}^2} \times \sqrt{\sum_{i=1}^{m} W_{ij}^2}}$$

2.8.Metrics for IR Evaluation

The aim of any Information Retrieval system is to search document in responce to a user query relavant to his information need. The performance of IR systems is evaluated on the basis of how relavent documents it retrieve.

Relevance depends upon a specific user's judgment. It is subjective in nature. The true relevance of the retrieved document can be judged by the user only, on the basis of his information need. For same query statement, the desired information need may differ from user to user.

Traditionally the evaluation of IR systems has been done on a set of queries and test document collections. For each test query a set of ranked relavant documents is created manually then the system result is cross checked by it.

There are many retrieval models/ algorithms/ systems. Different performance metrics are used to assess how effeciently an IR system retrieve the documents in responce to a users information need.

Different Criteria's for evaluation of an IR system are:
a. Coverage of the collection
b. Time lag
c. Presentation format
d. User effort
e. Precision
f. Recall

Effectiveness is the performance measure of any IR system which describes, how much the IR system satisfy a user's information need by retrieving relevant documents.

Aspects of effectiveness include:
a. Whether the retrieved documents are pertinent to the information need of the user.
b. Whether the retrieved documents are ranked according to the relevance with the user query.
c. Whether the IR system returns a reasonable number of relevant documents present in the corpus to the user etc.

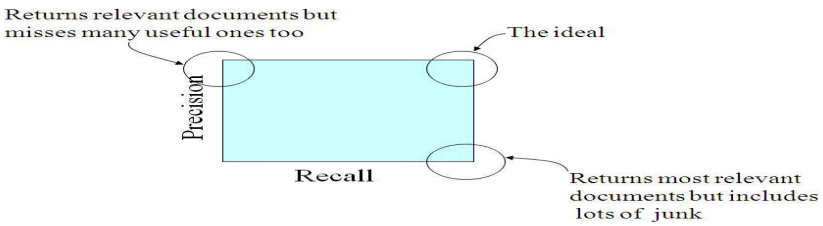

Fig. 3: trade-off between precision and recall

3 OUR APPROACH

In this work, to develop an Information Retrieval system for Urdu language, the following methods and evaluation parameters are used.

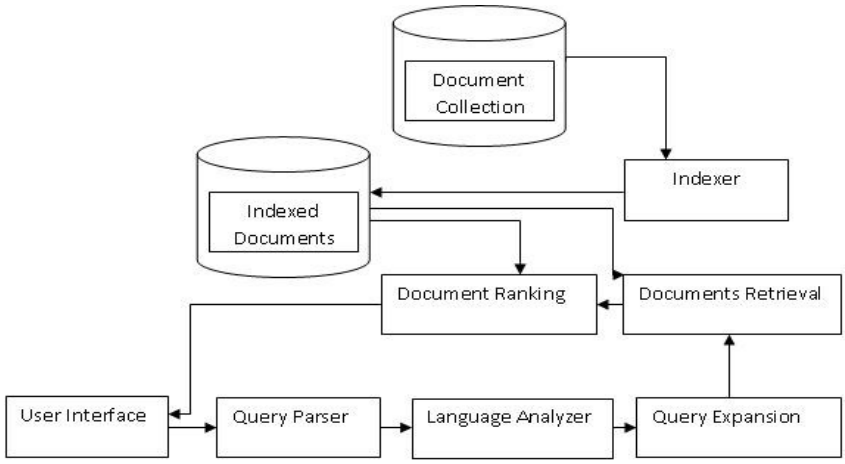

Fig. 4: Architecture of monolingual IR system

3.1 Stemmer:

For Developing Stemmer we have used an unsupervised approach [1] which gives accuracy of 84.2.

3.2 Term Weighting scheme:

For term weighting we have used the tf*idf

3.3 Indexing Scheme:

In this work the query statement and the documents are represented using the word indexing strategy.

3.4 Retrieval Model:

To implement our IR system we have used the vector space model.

3.5 Encoding Scheme:

As the system focuses on Urdu language, to access the data UTF8 character encoding is used.

3.6 Similarity Measure:

For getting documents which are more closely related to the query i.e. to measure the similarity between different documents in the corpus and the query statement, the cosine similarity measure is used.

3.7 Ranking of the document:

For ranking of the retrieved documents in order to their relevance with the query, cosine similarity values are used. The document having higher cosine value (min angular distance) with the query will be more similar i.e. contains the query terms more frequently and hence these documents will be considered more relevant to the user query.

4 EXPERIMENT

The data set used in this thesis for the training and testing of the developed Urdu IR system is taken from Emilie corpus. In this corpus documents are in xml format.

The data set taken from EMILLE corpus a tagged data set consist of documents having information related to health issues, road safety issues, education issues, legal social issues, social issues, housing issues etc.

The testing data set consist of documents from various domains such as:

Table 4: dataset specification used for Urdu IR

Domain	Number of Documents	Number of words
Health	33	223412
Education	8	115264
Housing	8	120327
Legal	8	108055
Social issues	12	146083
Homeopathy	32	527360
Drama	13	135680
Myths	10	202880
Story and Novel	21	300160
Media	15	224000
Science	47	704000
History	33	502400
Politics	21	728320
Psychology	27	555520
Religion	34	556800
Sociology	21	398080
Miscellaneous	48	985374

A Query set consist of 200 queries is prepared manually for training and testing of the IR system.

Table 5: Sample of Query set

بچوں کو پڑھنے میں دلچسپی کی اہمیت
کیٹرنگ مینیجر کے لازمی مسائل
روڈ سیفٹی
ان لوگوں کے لئے جو ریٹائر ہو رہے ہیں یا ریٹائر ہو گئے ہیں مالی مدد
سوشل سکیورٹی بینیفٹس یا ٹیکس کریڈٹس
سرویکل سمیئرٹیسٹ کیا ہے
سرویکل سمیئرٹیسٹ
اعضاء کا عطیہ کون دے سکتا ہے

5 RESULTS AND DISCUSSIONS

For testing purpose of the developed Information Retrieval system, a test collection of 350 documents have been used. A set of 200 queries was constructed on these 350 documents. This query set is used to evaluate the developed Urdu IR.

Table 5: results of the developed Urdu IR system testing

Number of documents	Number of queries	Precision		Recall	
		Min (avg.)	Max (avg.)	Min (avg.)	Max (avg.)
350	200	0.13	0.63	0.5	0.8

As shown in the above table, the system has value of 0.13 as the minimum average precision and maximum average precision value of the system is 0.63.

Similarly the minimum average recall value for the system is 0.5 and maximum average recall value was found out to be 0.8.

6 CONCLUSION AND FUTURE WORK

In this paper we have discussed various indexing schemes and IR models. We have used tf*idf scheme for indexing and to implement the IR system VSM (Vector Space Model) is used.
The experimental result shows that the average recall of the developed IR system is 0.8 with 0.3 precision.

IR is one of the hottest research fields. One can do a lot new research to provide efficient IR system which can satisfy the user's information needs. A lot of research is needed to develop language independent approaches to support IR systems for multilingual data collections.

REFERENCES

[1] Mohd Shahid Husain et. al. "A language Independent Approach to develop Urdu stemmer". Proceedings of the second International Conference on Advances in Computing and Information Technology. 2012.
[2] Rizvi, J et. al. "Modeling case marking system of Urdu-Hindi languages by using semantic information". Proceedings of the IEEE International Conference on Natural Language Processing and Knowledge Engineering (IEEE NLP-KE '05). 2005.
[3] Butt, M. King, T. "Non-Nominative Subjects in Urdu: A Computational Analysis". Proceedings of the International Symposium on Non-nominative Subjects, Tokyo, December, pp. 525-548, 2001.
[4] Chen, A. Gey, F. "Building and Arabic Stemmer for Information Retrieval". Proceedings of the Text Retrieval Conference, 47, 2002.
[5] R. Wicentowski. "Multilingual Noise-Robust Supervised Morphological Analysis using the Word Frame Model." In Proceedings of Seventh Meeting of the ACL Special Interest Group on Computational Phonology (SIGPHON), pp. 70-77, 2004.
[6] Rizvi, Hussain M. "Analysis, Design and Implementation of Urdu Morphological Analyzer". SCONEST, 1-7, 2005.
[7] Krovetz, R. "View Morphology as an Inference Process". In the Proceedings of 5th International Conference on Research and Development in Information Retrieval, 1993.
[8] Thabet, N. "Stemming the Qur'an". In the Proceedings of the Workshop on Computational Approaches to Arabic Script-based Languages, 2004.
[9] Paik, Pauri. "A Simple Stemmer for Inflectional Languages". FIRE 2008.
[10] Sharifloo, A.A., Shamsfard M. "A Bottom up Approach to Persian Stemming". IJCNLP, 2008
[11] Kumar, A. and Siddiqui, T. "An Unsupervised Hindi Stemmer with Heuristics Improvements". In Proceedings of the Second Workshop on Analytics for Noisy Unstructured Text Data, 2008.

[12] Kumar, M. S. and Murthy, K. N. "Corpus Based Statistical Approach for Stemming Telugu". Creation of Lexical Resources for Indian Language Computing and Processing (LRIL), C-DAC, Mumbai, India, 2007.
[13] Qurat-ul-Ain Akram, Asma Naseer, Sarmad Hussain. "Assas-Band, an Affix-Exception-List Based Urdu Stemmer". Proceedings of ACL-IJCNLP 2009.
[14] http://en.wikipedia.org/wiki/Urdu
[15] http://www.bbc.co.uk/languages/other/guide/urdu/steps.shtml
[16] .http://www.andaman.org/BOOK/reprints/weber/rep-weber.htm
[17] Natural Language processing and Information Retrieval by Tanveer Siddiqui, U S Tiwary.
[18] Information retrieval: data structure and algorithms by William B. Frakes, Ricardo Baeza-Yates.
[19] http://www.crulp.org/software/ling_resources.htm

AUTHOR

Mohd. Shahid Husain

M.Tech. from Indian Institute of Information Technology (IIIT-A), Allahabad with Intelligent System as specialization. Currently pursuing Ph.D. and working as assistant professor in the department of Information Technology, Integral University, Lucknow.