

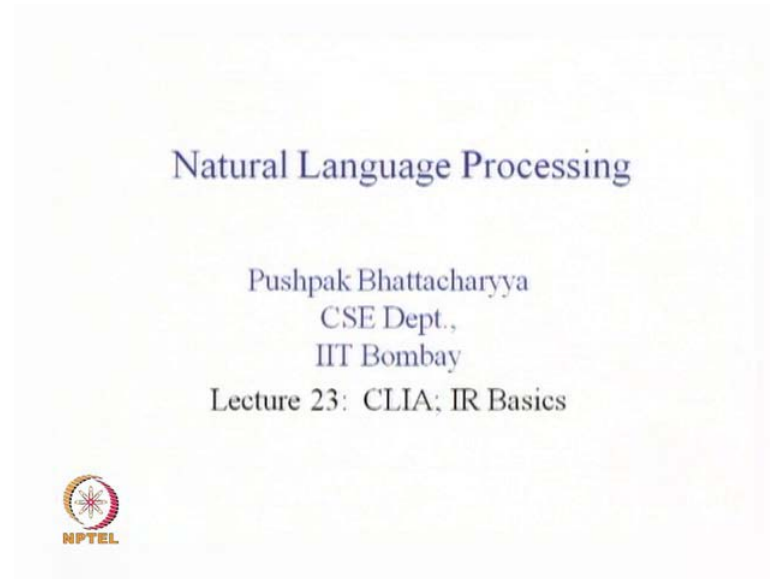
Natural Language Processing
Prof. Pushpak Bhattacharyya
Department of Computer Science and Engineering
Indian Institute of Technology, Bombay

Lecture No - 23
CLIA; IR Basics

In this course, now we have been exploring the relationships between two very large fields of computation artificial intelligence. These two fields are information retrieval and natural language processing. We remark that these two fields have a very natural connection between them because both the field deal with text primarily and are concerned with how the language phenomena are manifested in the form of: sentences, paragraphs, documents and so on. In information retrieval the key point is, the way the information need of the user is met. The user poses the query to the research engine and the documents which are retrieved are ranked according to the relevance of the document with respect to the users query.

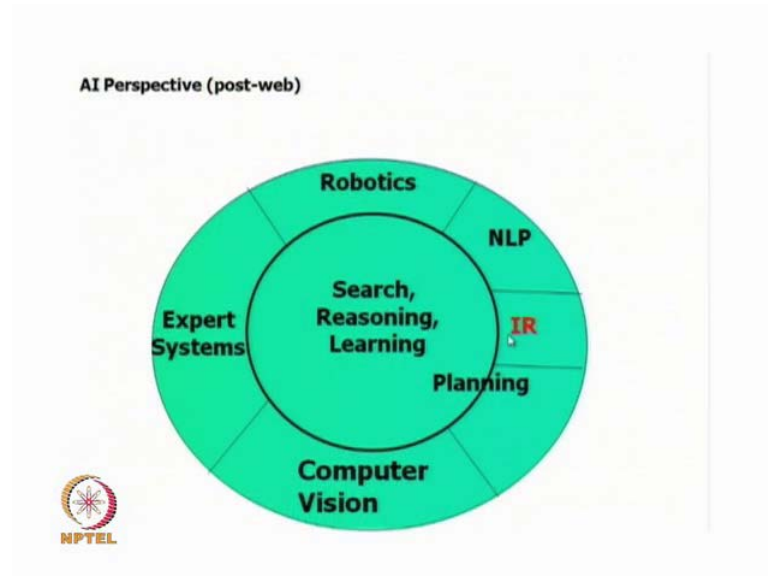
Natural Language Processing on the other hand, is concerned with the meaning content of the: documents, sentences, paragraphs and so on. And, clearly the meaning of a document has lot of bearing on the relevance of the document with respect to a query. We might remark here that there is a field of Natural Language Processing called summarization where the documents contents are summarized and presented to the reader. Now, the summary it can be formed in two different ways: one is query specific summary and the other is query independent summary. So, the user can have question in mind and the summary is presented to the user with respect to this particular question. The other kind of summary would be a general gist of a document independent of any query or question, all right. So, we understand that the information need is expressed in terms of query and the search retrieves information to satisfy that information need.

(Refer Slide Time: 02:47)



So, we proceed with a the lecture today and today's topic is describing cross linter information accesses which is a very large project in India and then move on to information retrieval basics. So, we started describing the cross liger information access project of the country which is a flashy project of the government. And, the reason is that the users have information need which is best expressed in the language of the user, the native tongue of the user. The comfort level with English in this country is about 5 percent of the population but the information need is a very real need and it is so happens that the information is expressed mostly in English. And therefore, we have to do something about crossing this language were all right.

(Refer Slide Time: 03:52)



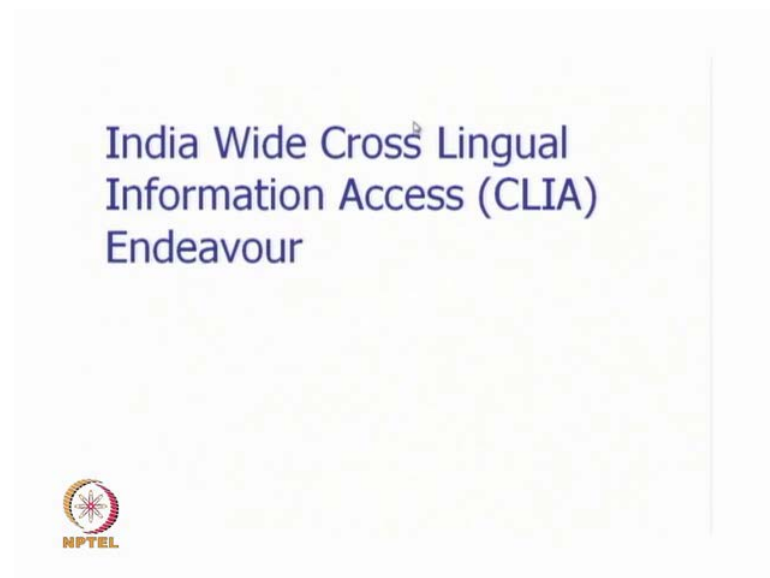
So, we proceed with the slides and I remind you that in Artificial Intelligence, information retrievals entry into the field is a recent phenomena because of the enormous present an impact and importance of the web. And, in the core areas of Artificial Intelligence namely: search, reasoning, learning, planning and so on. They do contribute to all these areas on the outer circle and information retrieval also makes use of A I techniques.

(Refer Slide Time: 04:32)



We remarked that the most important quantity in this discussion is user satisfaction with respect to his or her information need. And, this satisfaction depends on the way, the documents which are retrieved are ranked. According to their relevance this ranking on the other hand, is a function of the coverage of the documents. How much of the information from the web we have been able to capture and keep in our depository to satisfy the information need. The other point is that the user query needs to be processed correctly. We have to identify the: root words, the stem, the named entities or proper noun, multi words or compound words and so on. And, if we do all these properly then we get a set of documents which are relevant and the most relevant documents appear, does appear on the higher side of this list so and then user satisfaction is achieved.

(Refer Slide Time: 05:39)

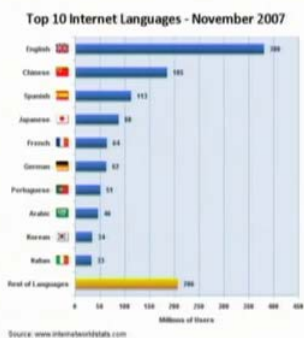


Now, in this context we started discussing India Wide Cross Lingual Information Access Endeavor and this is motivated by crossing the language areas between Indian languages and English. And also, amongst Indian languages so as to satisfy the information need of a user who know, let us say only one language namely his or her mother tongue or at least one more language.

(Refer Slide Time: 06:09)


Motivation

- English still the most dominant language on the web
 - ◆ Contributes 72% of the content
- Number of non-English users steadily rising all over the world
- English penetration in India
 - ◆ Estimated to be around 3-4%
 - ◆ Mostly the urban educated class
- Need to enable access to above information through local languages



Language	Millions of Users
English	380
Chinese	185
Spanish	113
Japanese	88
French	64
German	62
Portuguese	51
Arabic	46
Korean	34
Italian	33
Rest of Languages	206

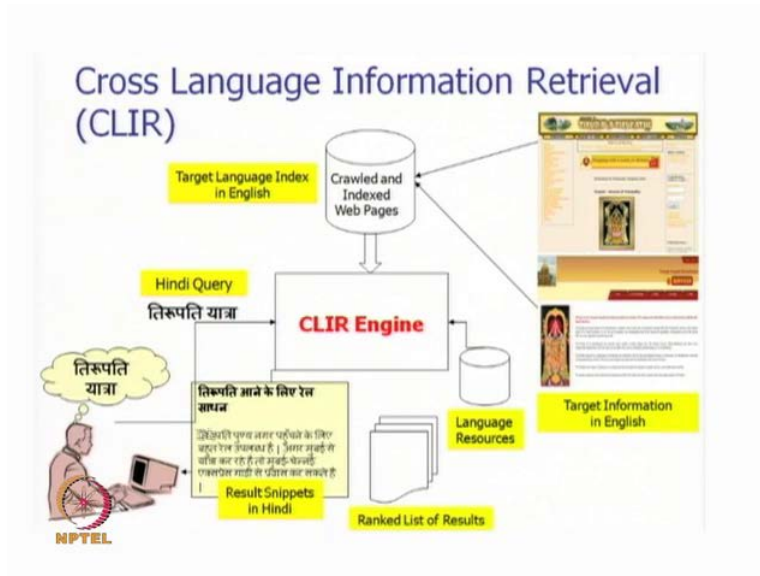
Source: www.internetstatistics.com
Copyright © 2008, Microsoft Marketing Group



So, English is still the most dominant language on the web that is the motivation for keeping English always in target. This contribute 70 percent of the content on the web. The number of non English users is steadily rising all over the world. That is also a fact as these view graphs or bar graph shows. We find that the top 10 internet languages as reported in November 2007, that is about 3 years back is English where 380 million users are English speakers. 380 million users of the web are English speakers followed by: Chinese at 185, Spanish at 113, Japanese 88, French 64, German 62, Portuguese 51, Arabic 46, Korean 34 and Italian 33 and rest of the language comprise about 206 million users.

But, this profile is changing quite rapidly, the number of non English users steadily rising all over the world. English penetrations in India is estimated to be around 3 to 4 percent mostly the urban educated class. And, we need to enable access to above information through local languages. So, this is a very real need which needs to be addressed.

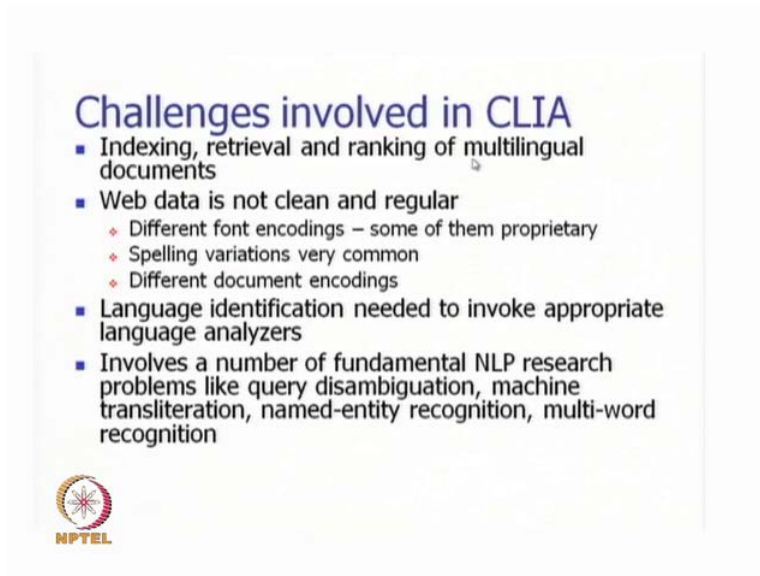
(Refer Slide Time: 07:38)



And, we showed this picture last time where the query here is shown to be Thirupathi yaatra that is, visit to Thirupathi. And, this is the query in the mind of the user. However, this is only an expression, partial expression of the real information need of the user who might to one two know many many things about Thirupathi. About how to reach Thirupathi, where to stay in Thirupathi, what are the hospitals in Thirupathi, where the police station is, how to go from Thirupathi to shrine of the lord, lord Venkatesh and which is Thirumala. So, all theses information needs to be fetched and shown to the user.


So, this query is given to the C L I R engine, the search engine which is then processed and converted into English by using language resources. And, this query is now taken to the crawled and indexed web pages. The index table matches the query word, the key words in the query with words present in the index table. And then, the documents are fetched which are on Thirupathi and there in English though and therefore, these have to be now shown in Hindi because Hindi was the language of the query. So, not the whole document need to be translated but only the important part of the document which is an indication of what it contain with respect to the query that is translated and then shown to the user. So, this is the process: Query, query processing, crawled web pages referral, fetching of documents and output shown to the user.

(Refer Slide Time: 09:35)



Challenges involved in CLIA

- Indexing, retrieval and ranking of multilingual documents
- Web data is not clean and regular
 - ◆ Different font encodings – some of them proprietary
 - ◆ Spelling variations very common
 - ◆ Different document encodings
- Language identification needed to invoke appropriate language analyzers
- Involves a number of fundamental NLP research problems like query disambiguation, machine transliteration, named-entity recognition, multi-word recognition



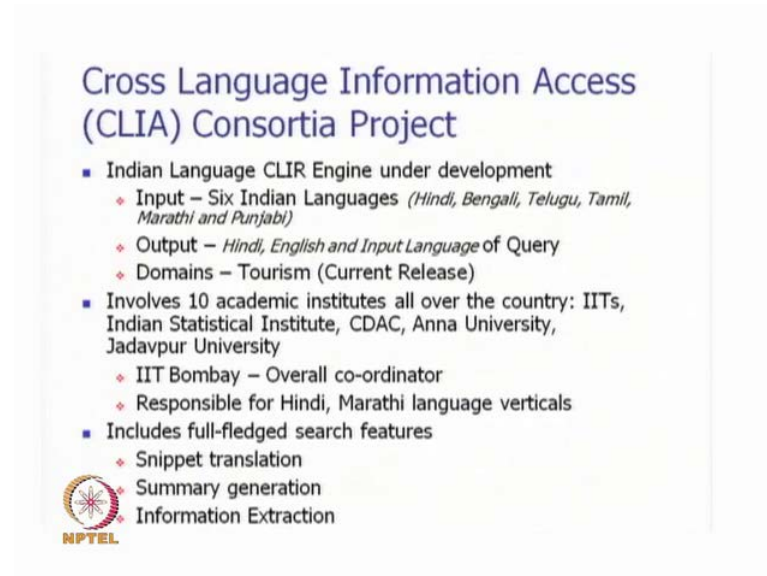
So, the challenges which are involved in Cross Lingual Information Access and well understand this better, when we do information retrieval basics. Indexing retrieval and ranking of multilingual documents is the main concern of the problem, indexing retrieval and ranking of multilingual documents. Web data is not clean and regular: different font encoding exists, some of them are proprietary, spelling variations is very common, different documents and encoding. So, encoding is indeed a very important concern all most head ache for people working with text. And, the reason is that the documents are many times is available in proprietary font.

For examples, the news papers have their own font typically in Hindi, Nava Bharath times, has its own font and another Hindi news paper also has its own font. And therefore, if any search train is used which is not in the same font, the encoded font of the newspaper it will not be able to match the content and retrieved the information. So, this is a very important problem and by the efforts of the Ministry of Information Technology and Communication there has been some amount of conscious set in form uniformity with respect to the encoding of Indian language font. Now, more or less people are used to adopting the u t f 8 or Unicode font to be used for all Indian language. So, this is a very important point because of this non uniformity and non standard font search translation, all these large activities do take a back seat unless we achieve these uniformity.

So, this is the point which was mentioned on the slides as the 2nd point, web data is not clean and regular. Language identification is needed to invoke appropriate language analyzers. This is also a very important point, in Devanagari for example, the languages which use Devanagari script are: Hindi, Marathi, Konkani, Nepali and Sanskrit. So, all these languages are used as same script namely Devanagari. Now, the question that arises is that if I look at a document on the web and we see that the document is in Devanagari. How do I know which language these documents are written in is it Hindi or is it Nepali or Marathi and so on. And, Hindi content of course, is very large on the web. Marathi follows Hindi but Marathi documents also are not very little. And therefore, when we retrieve documents with respect to a query, suppose the query is in Hindi the search in it should not retrieve Marathi documents.


So, language identification is very very important and if you are building language specific index tables to facilitate a particular language search then it is important to do this language identification. Then we see that processing of a query involves a number of fundamental NLP research problems like: query disambiguation, machine transliteration where by a proper noun is converted into the script of the target language, named entity recognition, multi-word recognition. Multi word recognition is important because for strings like good Friday this whole compound word has a specific meaning which is which pertains to Christianity and those documents have to be retrieved. Good Friday does not mean if Friday that is good, it is a specific meaning. So, these NLP tasks are involved.

(Refer Slide Time: 13:43)



Cross Language Information Access (CLIA) Consortia Project

- Indian Language CLIR Engine under development
 - ◆ Input – Six Indian Languages (*Hindi, Bengali, Telugu, Tamil, Marathi and Punjabi*)
 - ◆ Output – *Hindi, English and Input Language of Query*
 - ◆ Domains – Tourism (Current Release)
- Involves 10 academic institutes all over the country: IITs, Indian Statistical Institute, CDAC, Anna University, Jadavpur University
 - ◆ IIT Bombay – Overall co-ordinator
 - ◆ Responsible for Hindi, Marathi language verticals
- Includes full-fledged search features
 - ◆ Snippet translation
 - ◆ Summary generation
 - ◆ Information Extraction

 NPTEL

Now, the Cross Language Information Access project in the country is a consortia project. Indian language C L I R engine is under development. There are 6 Indian languages which are in the purview of the project. Recently 2 more languages have been added mainly Assamese and Gujarati. The output happens in Hindi, English and the input language of query. So, that means the output can be seen either in the language of the query which means monolingual information retrieval or in Hindi or in English.

The domain is tourism. It is proposed to add health and may be the agriculture. But, the addition of domain is sometimes a challenge because new resources have to be developed though the computational tools would work from one domain to another. And, this large project involves 10 academic institutes all over the country: I I T's, Indian Statistical Institute, C D A C, Anna university, Jadavpur university and so on. I I T Bombay overall coordinate and it is also responsible for Hindi Marathi languages verticals. Now, the whole system includes full fledged search features along with the: snippet translation, summary generation, information extraction. So, all these you understand is towards satisfying the information need of the user even if the user does not understand English.

Friday. Query translation achieve conversion of words into target language words. Query transliterations, achieves converting a proper nouns in the query language into the proper noun written in the script of target language. Query disambiguation produces the actual meaning of the words and their by improves the perception of the documents which are retrieved.

This query is then referred to the index table. The index table then produces the document which match the key words. So, searching is over now the documents which have come ranked in the order of their relevance, document summarization happens, snippet generation, template based information extraction happens. All these are translated and shown to the user in his own language. All this of course, does not come in free. They are supported by lots of offline processing.

So, the web is my the web is crawled for a documents. There are fonts trans coder, there are c m l i fier which embed reach tags in the documentary retrieval thrifts which are used for: output processing, information extraction, and so on. Language identifier identifies the language of the document. Domain identifier ascertains that the page is indeed from tourism domain which is our focus of work. Named entities are dictated and these are stored with special flags in this index table. This improves the ranking of the documents.

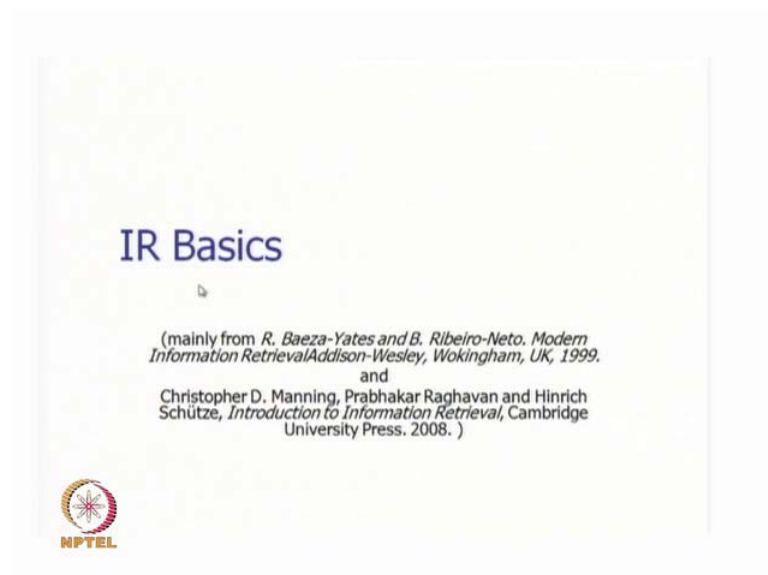
Multiword expression are also detected and stored with a flag mark this way in the index table. The information extraction data base is referred to for information extraction. There is another line of work in the whole system which is semantic search. There is a representation of sentences in the form of what is called universal networking language graphs. So, these search is also called semantic search. It is suppose to produce more high quality results and the user interface also bears the burden of taking user query and present the information in a lucid attractive way to the user alright. So, this is the top level block diagram of Cross Lingual Information Access this project which is going on in the country under the leadership of IIT, Bombay. So, these stars show various activities which are under different stages of development: crucial activities with different colors, not so crucial activity also are shown.

(Refer Slide Time: 18:54)




Now, this project has got quite a lot of coverage in the newspaper, leading dailies of covered our work in terms of what the impact of this project will be alright.

(Refer Slide Time: 19:10)



Now, we move on to information retrieval basic and we have made use of two well known text: one is modern information retrieval by Recirdo Beaza Yates and b Ribeiro Neto and the other book is more innocent one which is the introduction to information retrieval by Manning Raghahvan and Schutze. So, these text books provide us with some of the basic material which are important to have a good grip on a information retrieval.

(Refer Slide Time: 19:41)



Motivation

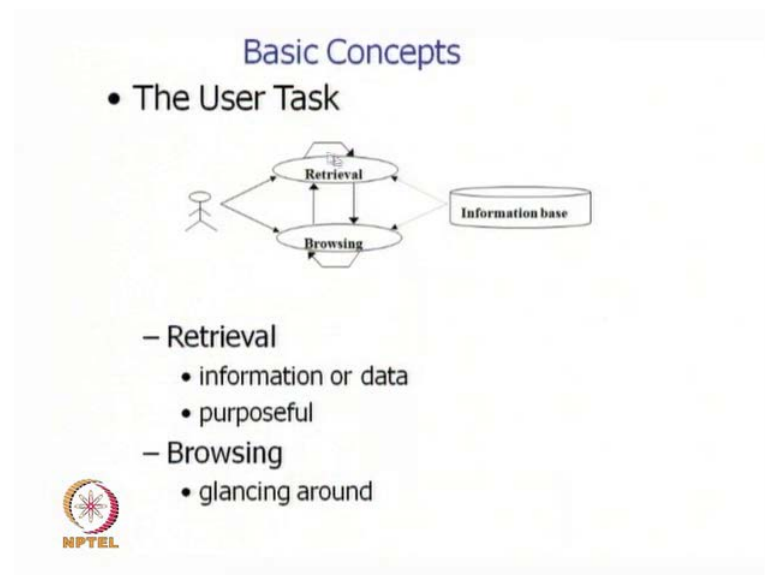
- IR at the center of the stage
 - IR in the last 20 years:
 - classification and categorization
 - systems and languages
 - user interfaces and visualization
 - Still, area was seen as of narrow interest
 - Advent of the Web changed this perception once and for all
 - universal repository of knowledge
 - free (low cost) universal access
 - no central editorial board
 - many problems though: IR seen as key to finding the solutions!

So, the reason for studying I R as a basic course at various academic institutes, search lab etcetera is that I R is now really at the center of the stage of the activities. In the last 20 years I R has been mainly concerned with: classification and categorization of information's systems and languages, user interfaces and visualization. But, I R was limited in its scope and people's interest in I R also was in specific problem, specific domains and so on. For example, in the library information system retrieving information about books was considered an important part of information retrieval. And, the book search could happen in terms of authors, that book titles and sometimes even on the topic in which the book is, the topic of book. So, in library information system was an important consumer of information retrieval techniques.

So, however the area was still seen as a area of narrow interest. Advent of the web changed this perception once and for all. Web has become such a granitic repository of information with the need for retrieving information in a: sophisticated way, a high quality access, a fast access. These things have changed the free low information retrieval completely. So, web is looked up as a universal repository of knowledge. It is almost like a fairy tale figure the wise man, the universal wise man who knows everything. So, people go to the web first for any kind of information need; it is universal repository of knowledge. The information is free or very low cost with universal access. There is no moderation which is its strength with little bit of weakness but that indeed is that strength.

Everybody is contributing information, unhindered inhibited to the web and that is making it more and more powerful. There is no central editorial board. Nobody is looking on anybody; nobody is looking over the shoulder of anybody to say that yes this information you should not put on the web, this information you should not look at. This is completely free territory of information which is a tremendous power as far as information access is concerned. And, many problems do exists many problems is, do exists because information retrieval does not happen free of course. There are many, many challenges, large challenges to be solved. However, this is a new world, a free world with information for all and this is a tremendous achievement of computing networks, a Artificial Intelligence, language processing and so on alright.

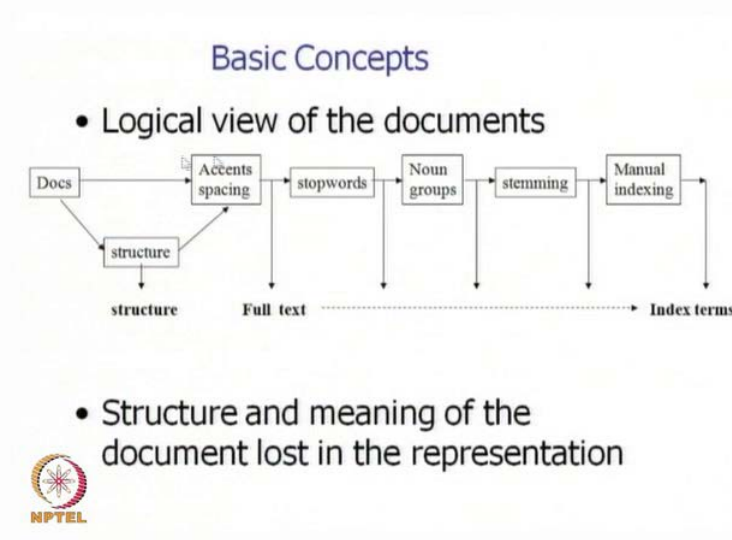
(Refer Slide Time: 22:59)



So the, we go to the slide and see that the user task and information retrieval can be modeled by means of this semantic. So, user pose a query to the retrieval system, the information base is referred to what comes out is presented to the user. The user can do browsing of the web. A retrieval is information or data which is retrieved. Retrieval is purposeful. Browsing on the other hand may not be completely purposeful, is simply glancing around whatever is there on the web. So, this situation can be likened to a tourist who visits a place with a particular purpose in mind. For example, I go to Jaipur with a specific purpose of seeing the city palace.

So, this would be keen to purposeful retrieval. Another situation is I go to Jaipur and simply roam around looking at interesting things and looking at also things which are ordinary. But, still found part of the life and sights and sounds of the city. So, the difference between browsing and retrieval is this.

(Refer Slide Time: 24:25)



The basic concepts can be also captured by means of these pipeline or a semantic which gives the logical view of the documents. So, the documents have their own structure. There are: accents, spacing etcetera. Paragraphing, listening, chapters these are the structure specific features of the document. So, when were these: stags, paragraphs etcetera are obliterated then we have a large string which is corresponding to the full text of the document. There are stop words, let me in words which are present in all document and therefore, they do not length any special information with respect to the identity or specific information content or document.

So, the stop words are those entities which are required to form the document but are not distinctive feature of the document. An example of this would be the articles like: a, n, a, an, the. They are presented all document and therefore, they do not bringing any specific information about a specific document. When information retrieval happens it is the nouns and noun groups which play the central role in information retrieval. They are: nouns are the most important carriers of the information, verbs, preposition, adjective, adverbs etcetera, serve either two qualified this nouns are to glow together the nouns to

found a particular information content in the noun group past through what is called stemming.

Stemming is very crucial because the form of a word in the query may not be the same as the form of the word in the document that satisfied the information. For example, the word docs is the plural form of doc and the single form is doc. The query may contain the word docs in a single form. The document may contain it in its plural form and then if the system says that these two entities, namely: the query and the document, do not match because there is a singular and plural difference, singular and plural disparity. Then this is not the right situation. The document does not contain the information to satisfy the query just because the word is not in plural form. This is not a good idea to say that this document is not relevant.

So, words typically transform themselves according to morphological information, according to features and when that happens it is important to bring them up to a common step. Docs and doc should both be docs and then the matching can take place. So, this is the importance of stemming. There are languages which have very complex morphological processes. The word gets transformed in a complicated way and in a rich way into many different forms. So, the query and document do not match because of a morphological variation, it's not a good idea. So, stemming is very crucial.

Then we find that there is manual indexing. Manual indexing takes the stems and essentially creates a very large array of words which point to the documents that contain them. This is also called inverted indexing. We are going to discuss all this in detail. Now, when the query processing happens, stemming happens, paragraph tags etcetera are removed. Then we can see that the structure and meaning of the document is lost in the representation. So, this is a very important point to note. The document comes with its structure and meaning. When it is gone through this kind of: this structuring, stop word removal, identification of nouns and they are consequently stemmed by the suffixes script then what is happening? Then maybe we should say that the document has been scripted down to its essential elements.

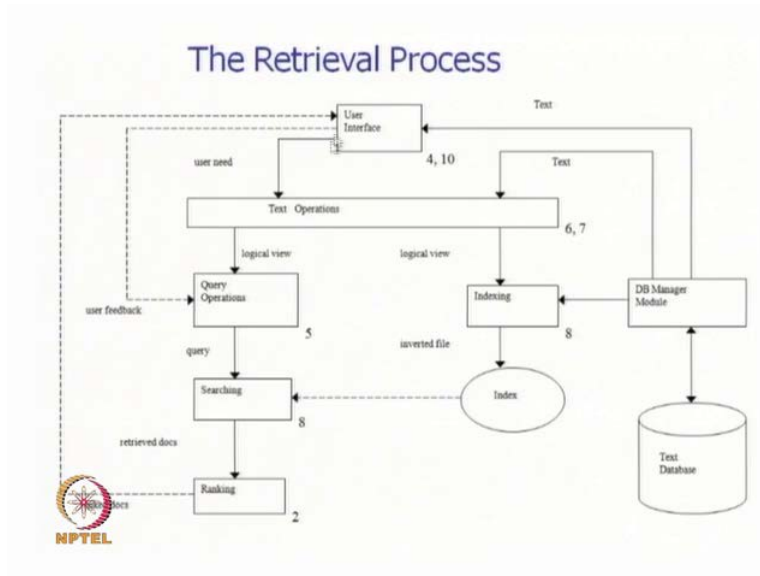
The elements which are required for presenting to a user a logical view of the document, that logical view is simply saying that these are the important keywords in the document. And, whatever with the information need if this information need, it contains those

keywords. Then these document would satisfied that information needs. It is very likely to satisfy the information needs but the fact remains that the document representation to the search engine is in terms of a set of keywords organized in a table. And, that is hardly close to whatever the information, whatever meaning the document contents. So, these an important point to keep in mind. If you compare this with another field of Natural Language Processing namely machine translation then it is crucial in machine translation to keep the meaning content in tacked.

And, in that meaning content the stop words and article preparation, conjunction they play very crucial role. So, we sought of tent to see that things which are important in Natural Language Processing and information retrieval have some community and some different and important. So, nouns are important in both places: information retrieval, N L P. Both places have given lot of importance to nouns. They are the carries of information. Information retrieval does not care much about: stop word, articles and preparation conjunction etcetera. But, in machine translation an N L P these words play a very important role. They cannot be deleted, they cannot big note.

So, this we see as an important difference between Natural Language Processing and information retrieval. So, they coincide in terms of nouns but, they diverge in there is important given to the function words. Let us keep this point in mind, content word important for both, function word important for N L P, crucial for N L P but not so for information retrieval ok.

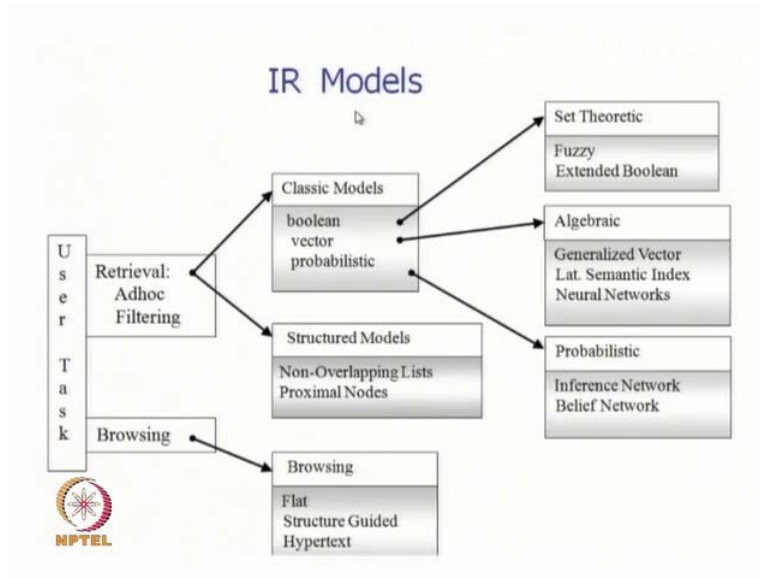
(Refer Slide Time: 31:04)



So, proceeding with the material, the retrieval process happens in this way. The user interface takes in the query. There are, so the query is a an expression of the user information need. The text operations takes place on the query. There is a logical view of the text operations that means, essentially keywords. The query operations happens and it is possible at this stage to get the user feedback, we will see it later as to how it happens. Then the query operations launches search. The search retrieves documents which are ranked and shown to the user. The user based on the rank documents can give a feedback which can come to the query box again and then again the search process would begin. So, these looping can happen until the user is satisfied with the retrieval.

In practice, whether this happen is not, this is a separate question; we will discuss this later. And, in the offline at the background something else this happening which is that data base manager that has a massive text database at disposal. Those pages, those documents are indexed, which again is a logical view of the information repository, the huge information repository which needs to be referred to frequently. And then, there is inverted file which means that the there is an indexing created and it is called inverted because, we have mapping from words to documents, rather than the other way. So, word to document is the inverted index.

(Refer Slide Time: 33:03)



Now, we come to a very deep and important point namely: the issue of information retrieval models, IR models. So, user task is retrieval, which can be in terms of Adhoc filtering or the user task is to do browsing. So, when browsing happens there is this flat files system, structure guided file system or hypertext. So, one could brows either a flat file which could be a: dot text, dot p d f, dot doc any of those files or it could be semi structure data or a completely structure data. Hypertext is semi structure data with tags embedded in text and completely structured data would be tables for examples. Under structure models, we have this non overlapping lists.

There are these proximal nodes where, the data base in the form of entity and the relationships is a large graph of the context of the data. And then, there are these classic models information retrieval namely: Boolean, vector and probabilistic models. So, what are we saying here? What we are saying here is that the information retrieval model tries to capture relevance. The most important concern of information retrieval model is trying to capture this elusive concept of relevance. How does the one capture relevance? That means how do you know that the particular query, the whatever is contained in the document is relevant to that query? This is a question of relevance , very important point.

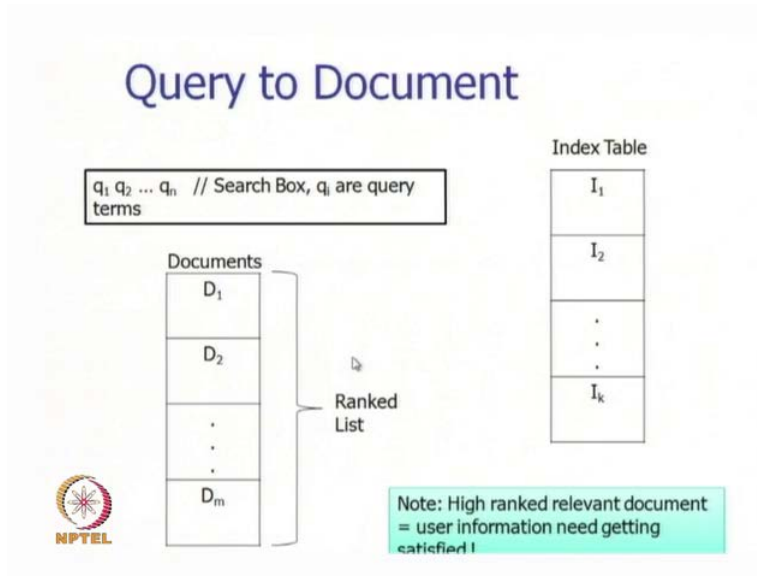
So, the slide says that the classic models do relevance modeling which is same as IR modeling either through Boolean operations or vector space operation or probabilistic

modeling. We will discuss this one by one and as we do so, we also bring in how NLP can or can't play a role in this very crucial concern of information retrieval namely: relevance. Boolean vector and probabilistic models of IR or relevance referred to: set theoretical modeling, fuzzy modeling, extended Boolean modeling. Algebraic: generalized vector, latent semantic index and neural networks. Probabilistic is: inference network, belief networks etc. So, under set theoretic we have: Fuzzy and extended Boolean modeling, under algebraic we have: generalized vector, latent semantic index based modeling, neural networks modeling, under probabilistic we have: inference network and belief network based modeling.

So, I would like to once again point that these are different terms and one should not lose sight by the force of terminology of the fact that the sole purpose is to capture relevance. How do we match the query and document? What is the best way to match them? So, this matching can happen at various levels. It can happen at a very superficial level where, the surface entity is on the query and the surface entities on the document are matched. On the other hand, the matching can be much more differed, the meaning of the query and the meaning of the document are matched and then the relevance is deserved.

So, it is clear that if matching at deep level is conducted; carried out then we have likely to far more relevant and far more accurate effective model of relevance. But, this may not be practical at the scale of the web where, the user's patience with the search engine can be only so much etc. If I am not receiving an answer to my query, however imperfect in about a second then I will use patience with the search engine, however sophisticated its operations are etc. So, see the psychology of a search engine user. A search engine user can tolerate some imperfectness. It can tolerate some limitations in accuracy so on but the speed of the response has to be really fast and then the user will think I would apply my mind and look at the list of document and get by information need satisfied etc.

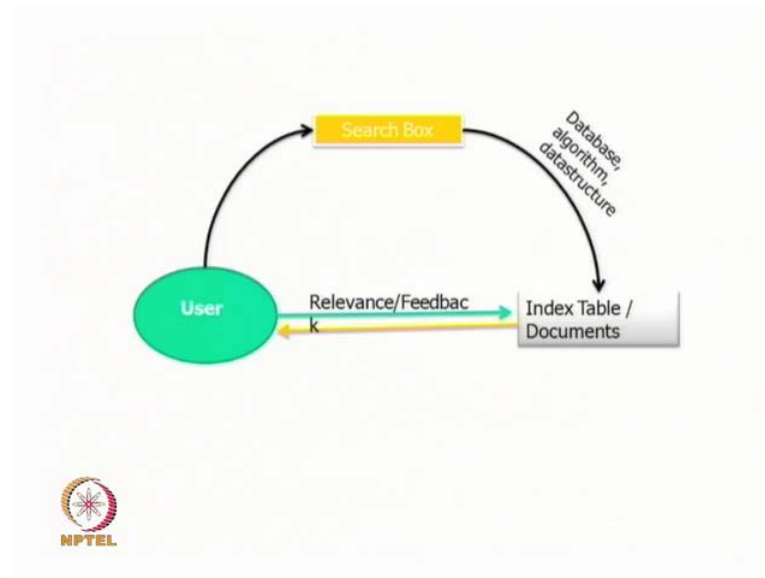
(Refer Slide Time: 37:48)



So, we proceed further and look at this very important slide which captures the essence of information retrieval. We have here, the query words given q_1, q_2, q_n , the search box q_i are query terms. The documents here are represented as a ranked list with respect to the query. And, how the document came; how it came was through this index table. This index table contains the terms which, point to least of document with respect to each term. For example, the word could be Taj mahal and all the document which contain with Taj mahal are link to this serve of the index table ok, in the link list fashion. So, if the thousands of document which contain the Taj mahal then, the Taj mahal entry in the index table will to contain a pointer to all these documents.

The query words are taken. Their stripped to their stems so that, the stems present in the index matched. So, the query a keywords are than matched against the variable matching occurs we pick up the document and bring it for relevance calculation and ranking. So, if high ranked relevant documents are present, relevant documents are given high rank so that they in the top of the list then, user information need getting satisfied. So, this diagram would tell us what is the best way, ok. We could actual match documents, bring them and then order them according to their relevance. So, the rank list is the final product of searching and that would require language analysis in a, when you go to more sophisticated level.

(Refer Slide Time: 39:50)



So, here is a diagram which also incorporate what is called relevance feedback. So, user has represented his query in the search box. Then database algorithm, data structures all of them coming to picture. The index table and documents are refer to, these are shown to the user and relevance feedback can be given to the search engine once again saying that, I found the 5th relevant document to the very relevant. So, consider using the terms present in this document added to the query and the search is relaunched. This time the retrieval performance is better. So, it is a closed loop situation.

(Refer Slide Time: 40:33)

How to check quality of retrieval (P, R, F)

- Three parameters
 - Precision $P = |A \cap O| / |O|$
 - Recall $R = |A \cap O| / |A|$
 - F-score = $2PR / (P+R)$
 - Harmonic mean

The Venn diagram shows two overlapping circles. The left circle is labeled 'Actual(A)' and the right circle is labeled 'Obtained(O)'. The intersection of the two circles is labeled 'A ^ O'.

above formula are very general. We haven't considered that the documents retrieved are ranked and thus the above expressions need to be

Now, we come to very important point which is how to check the quality of retrieval. This is done in terms of precision recall and what is called the f score, we will explain all this carefully. So, we make some remarks with respect to the measurement of quality of retrieval; quality of the search engine. So, the point is that what is written from the engine is list of document. How good is this list of documents? Do you have a lot of relevant document in this list and are the relevant document appropriate placed? So, that an irrelevant document is towards the end of the list, relevant document is towards the beginning of the list.

All these are question with respect to the quality of the retrieval. So, these questions are answered by means of some objective parameters namely: precision, recall and f score; we now proceed to define them. We define them by means of sets but, this can be adapted to list situation where the document which come from the web are formed in to a list. So, the slide would show very precise definition of: precision, recall and f score. Look at these two circles so, here is this actual set of things denoted by the symbol A and what we have obtained is a set of things we call it the set O, A and O.

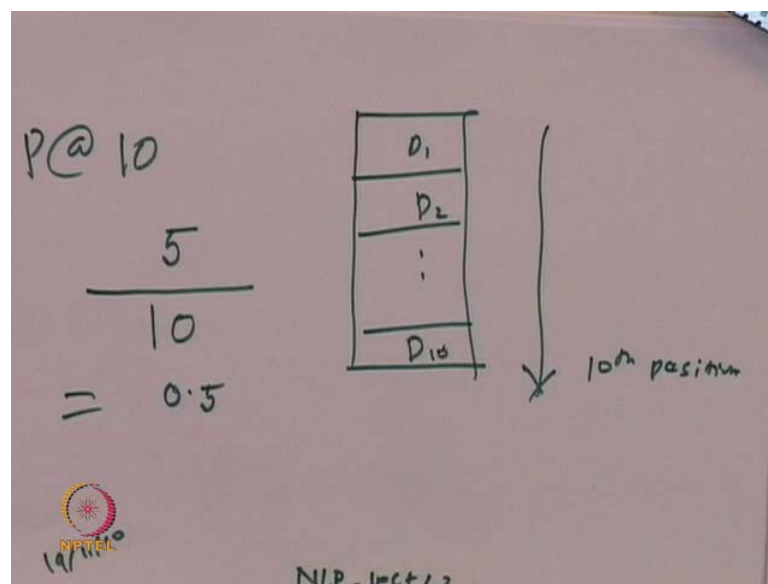
Now, it is conceivable that A and O, the actual set of things and obtained set of things intersects and produce this region. This is called the region of intersection between A and O. So, this is the region where adapted the documents or obtained entities or the actual entities agree. How is precision measure? Precision is measure as A intersection O. The size of A intersection O divided by the size of O. So, the agreement, the set of agreed upon entities: actual entities, the obtained entities divided by size of the obtained entities. So, precision is actually measuring if you look at the formula, it is actually measuring proportion of entities which are obtained and are correct.

It is measuring the proportion of correct entities out of the obtained once. This is the ratio, this is clear. So, out of the once which are obtained, how many are correct? These are proportion and if you multiplied by 100 then you will get the presentation precision. Recall on the other hand, is the capability of the system to actually obtained things which are actually correct. So, this expressed as the ratio of A intersection O, this area: size of this area divided by the size of A, the actual entities. So, this is giving me a proportion of the entity is from the actual set that are retrieved. So, do carefully note the definition of precision and recall. Both of them have the same numerator ok.

And, if we divide the area of intersection divided by the size of what is obtain than we get precision which essentially refers to this point of amongst everything that is obtained, how many are correct. Recall, on the other refers to amongst everything that is actually present, how much is retrieved? So, this is precision, this recall. There is a measure which combine precision and recall that is called f score. This is essentially the harmonic mean of precision and recall, $2 p r$ by p plus r . If the question is why is it harmonic mean and why not arithmetic mean or geometric mean? Then, just make observation that harmonic mean is the least amongst all this means. Harmonic mean is less then geometric mean, less than arithmetic mean.

So, if harmonic mean improves automatically geometric mean and arithmetic mean also will improve. Now, the other thing one would note is that if A and O are same then, precision recall and f score all boil down to a single number. So, one you possibly recall that in part of speech of speech tagging, we also used precision and recall. So, precision, recall and f score as inner terms which can be used for the measurement of agreement between an actual set of entities and the obtained set of entities. So, we can of course, make use of this in our information retrieval situation also where, we find that the list of the document which are present. So, let us say we look at the least up to the 10th position. If you look at the least of the 10th position I will draw this on the paper.

(Refer Slide Time: 46:11)



So, the documents which are obtained: D 1, D 2, up to D 10, so we go up to the 10th position. Then we might ask, what is the precision up to the 10th position which is written as P at 10? So, what is the precision up to the rank of 10? So, this can be easily answered. We simply find out of the 10 documents, out of these 10 documents how many are correct? So suppose, 5 are correct, 5 are relevant. Then $5 \div 10$ or 0.5 is the precision at 10. So, that is quite easy to calculate. Now, we come to an important point. Precision for information retrieval is defined and practically calculated. What can we say about the recall? Now recall, we say is the proportion of actual entities which are obtained.

So, the a particular process has obtained a set of entities. How many of them are what proportion of these form a component of what is actually existing? Now, here you can see the difficulty of the problem. It is difficult to answer the question easily. Why so, because nobody knows, what is the actual amount of information on the web. How many documents are relevant to a query? Nobody knows, where be as notion, very, very large repository of information. So, with respect to a query if we find out what is the recall performance of search engine, it is impossible to answer this question.

So, suppose we concentrate on, only on the 1st 10 document and out of them 5 have been found relevant. How this relevance judgment is we will see later. But, this 5 forms what proportion of the actual document present in the web, nobody knows. So, that is why the recall is not directly calculated for measuring the search engine performance. We will see what we do about this. Mainly what is calculated is precision at a particular rank and these values are computed different ranks we take an average of this rank. So, that is called mean average precision. We will discuss these issues in the next lecture.