

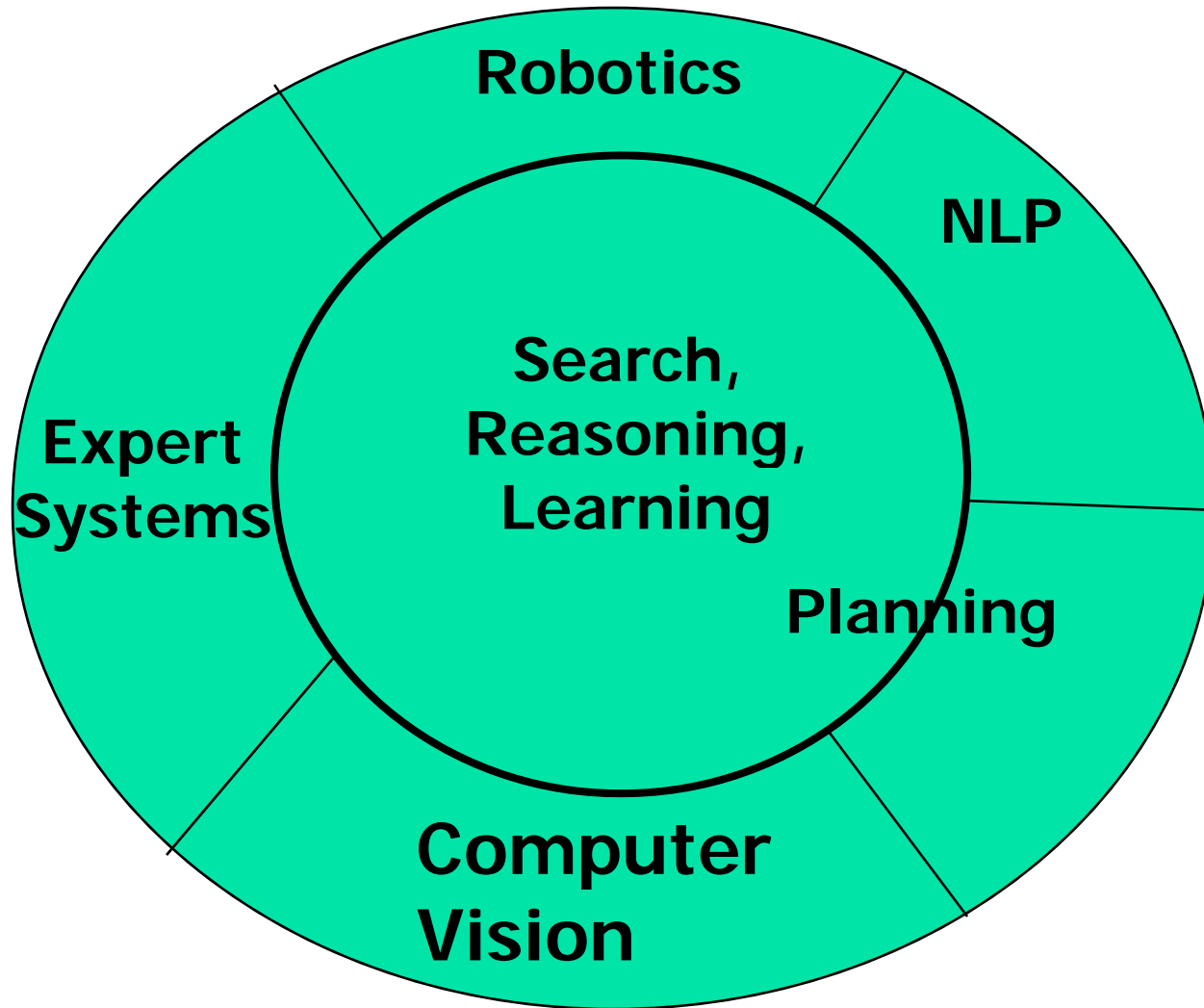
CS344: Introduction to Artificial Intelligence

Pushpak Bhattacharyya
CSE Dept.,
IIT Bombay

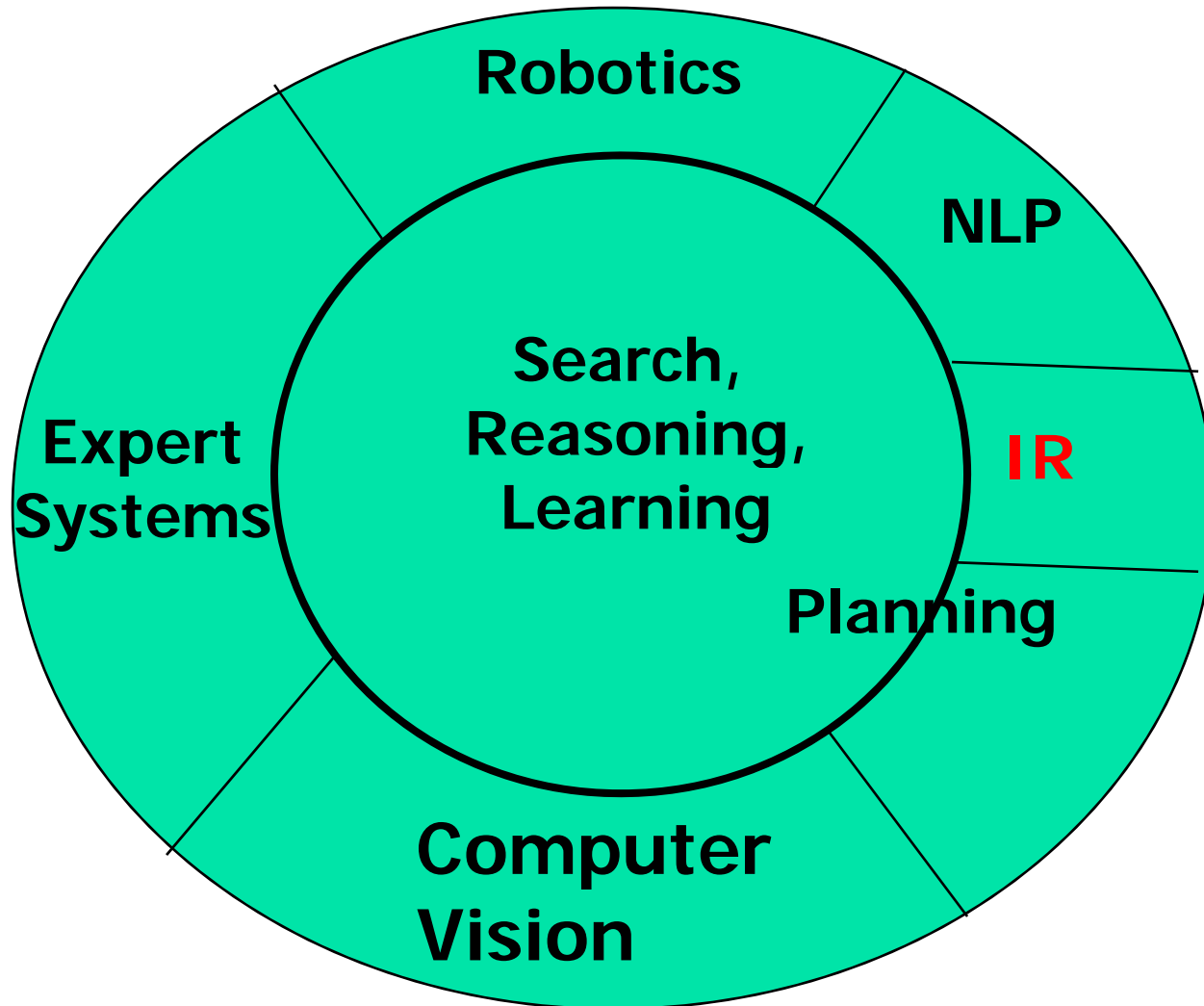
Lecture 31: Information Retrieval

IR and AI

AI Perspective (pre-web)

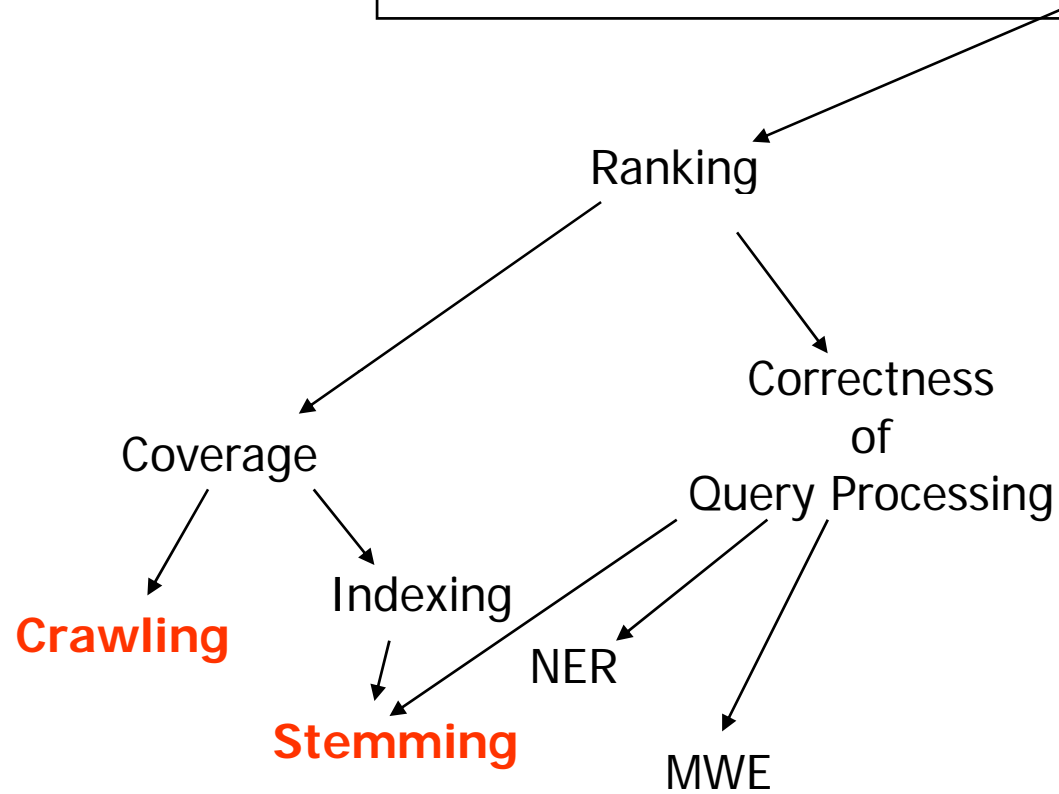


AI Perspective (post-web)



The elusive

user satisfaction



Query: Indian Tribes in Latin
America

Google

- **[Indians of Latin America: an exhibition of materials in the Lilly ...](#)**
- Lilly Library: *Latin American* mss. Brazil. A large map in colors, this locates the course of rivers, towns, mountain ranges, and *Indian tribes*. ...
www.indiana.edu/~liblilly/etexts/ila/ - 241k - [Cached](#) - [Similar pages](#) - [Note this](#)
- **[Indigenous peoples of the Americas - Wikipedia, the free encyclopedia](#)**
- *American Indian* creation legends tell of a variety of originations of that it had confirmed the presence of 67 different uncontacted *tribes* in Brazil, ...
en.wikipedia.org/wiki/Indigenous_peoples_of_the_Americas - 178k - [Cached](#) - [Similar pages](#) - [Note this](#)
- **[Cognition :: Giving Technologies New Meaning](#)**
- The volumes that Farabee produced from his travels include *Indian Tribes* of Eastern Peru ... motor vehicles that are lemons · *Indian tribes of Latin America* ...
wikipedia.cognition.com/?num=10&from_val...Indian%20tribes%20of%20Latin%20Ame... - 54k - [Cached](#) - [Similar pages](#) - [Note this](#)
- **[Top 25 American Indian Tribes for the United](#)**
- Top 25 *American Indian Tribes* for the UnitedStates: 1990 and 1980--Con. ... 16028 73.0 Canadian and *Latin American*... 19375 248.3 Chickasaw. ...
www.census.gov/population/socdemo/race/indian/ailang1.txt - 6k - [Cached](#) - [Similar pages](#) - [Note this](#)
- **[Ten Largest American Indian Tribes, 2000 — Infoplease.com](#)**
- *Latin American Indian*, 180940. Choctaw, 158774. Sioux, 153360 ... *American Indian* and Alaska Native Population by Selected *Tribes*, Census 2000 ...
www.infoplease.com/ipa/A0767349.html - 29k - [Cached](#) - [Similar pages](#) - [Note this](#)
- **[The Indian Tribes of North America by John R. Swanton at Questia ...](#)**
- Read the complete book *The Indian Tribes of North America* by becoming a Sao Paulo recently elected its...must cope with demands by *Latin America* for ...
www.questia.com/library/book/the-indian-tribes-of-north-america-by-john-r-swanton.jsp - [Similar pages](#) - [Note this](#)

Yahoo

- [different indian tribes of latin america](#),
- [More...](#)
- **WEB RESULTS**
- [South America Daily](#)
- **Indian** Pepper Photos Prices Spices. The Times of India ... Archaeologists unearth ancient **tribe** members sacri London ... Iran and the left in **Latin America** ...
- [www.wn.com/LatinAmerica](#) - 192k
- [Native American Indian Cultures - Mexico, South America](#)
- Also, many of the Yanomamo **tribe** are losing their members and culture by ... of Amazon **Indian** tribal art in the world, with over 75 **tribes** represented. ...
- [indian-cultures.com](#) - [Cached](#)
- [Native American Indian Cultures - links](#)
- North American **Tribes**. rednation.org - RedNation of the Cherokee. Meso and **Latin American Indians** ... Human Rights in **Latin America** ...
- [www.indian-cultures.com/Cultures/Links.html](#) - [Cached](#)
- [Indigenous peoples of the Americas - Wikipedia, the free encyclopedia](#)
- ... in **America**, particularly with regards to native **Indians**. ... Uncontacted **Indian tribe** found in Brazil's Amazon. The Peopling of the American Continents ...
- [en.wikipedia.org/wiki/Indigenous_peoples_of_the_Americas](#) - 179k - [Cached](#)
- [Native American Images - American Indian North America Tribe Map](#)
- American **Indian North America Tribe** Map. Click here to view more images ... Medal | History Hotline | Iraqi War | Korean War | **Latin Americans** | Medal of ...
- [www.nativeamericans.com/NativeAmericanImages6.htm](#) - [Cached](#)
- [Resources for](#)
- Numbers of Native Americans or **Indians** in **Latin America**: 39,442,000 million ... **Indian Tribes** in **Latin America** - **Latin American Indian** Population - Up date ...
- [www.xmission.com/~amauta/population.htm](#) - [Cached](#)
- [Indian tribe found in Brazil's Amazon - Boston.com](#)
- **Latin America/Caribbean**. **Indian tribe** found in Brazil's Amazon ... Uncontacted **tribes** are usually discovered when loggers and ranchers encroach on ...
- [boston.com/news/world/latinamerica/articles/2007/06/01/.../](#) News

AltaVista

- [Latin America](#)
Compare airfare prices from over 120 top websites and save up to 70%.
Flights.SideStep.com

[Regional Telecom Statistics & Forecasts](#)
Fixed, mobile, Internet, broadband telecom statistics and forecasts.
www.hottelecom.com
- [AltaVista found 4,520,000 results](#)
 - [South America Daily](#)
Indian Pepper Photos Prices Spices. The Times of India ... Archaeologists unearth ancient **tribe** members sacri London ... Iran and the left in **Latin America**
...
www.wn.com/LatinAmerica
[More pages from wn.com](#)
 - [Native American Indian Cultures - Mexico, South America](#)
Also, many of the Yanomamo **tribe** are losing their members and culture by ... of Amazon **Indian** tribal art in the world, with over 75 **tribes** represented. ...
indian-cultures.com
[More pages from indian-cultures.com](#)
 - [Indian tribes in Suriname cross borders - Boston.com](#)
Days of rain near Suriname's southern border have deluged Amerindian farmland, ... **Latin America**/Caribbean. **Indian tribes** in Suriname cross borders ...
www.boston.com/news/world/latinamerica/articles/2006/05/12...in_suriname_cross_borders
[More pages from boston.com](#)
 - [Indigenous peoples of the Americas - Wikipedia, the free encyclopedia](#)
... in **America**, particularly with regards to native **Indians**. ... Uncontacted **Indian tribe** found in Brazil's Amazon. The Peopling of the American Continents ...
en.wikipedia.org/wiki/Indigenous_peoples_of_the_Americas
[More pages from en.wikipedia.org](#)
 - [Native American Images - American Indian North America Tribe Map](#)
American **Indian North America Tribe** Map. Click here to view more images ... Medal | History Hotline | Iraqi War | Korean War | **Latin** Americans | Medal of ...
www.nativeamericans.com/NativeAmericanImages6.htm
[More pages from nativeamericans.com](#)

MSN

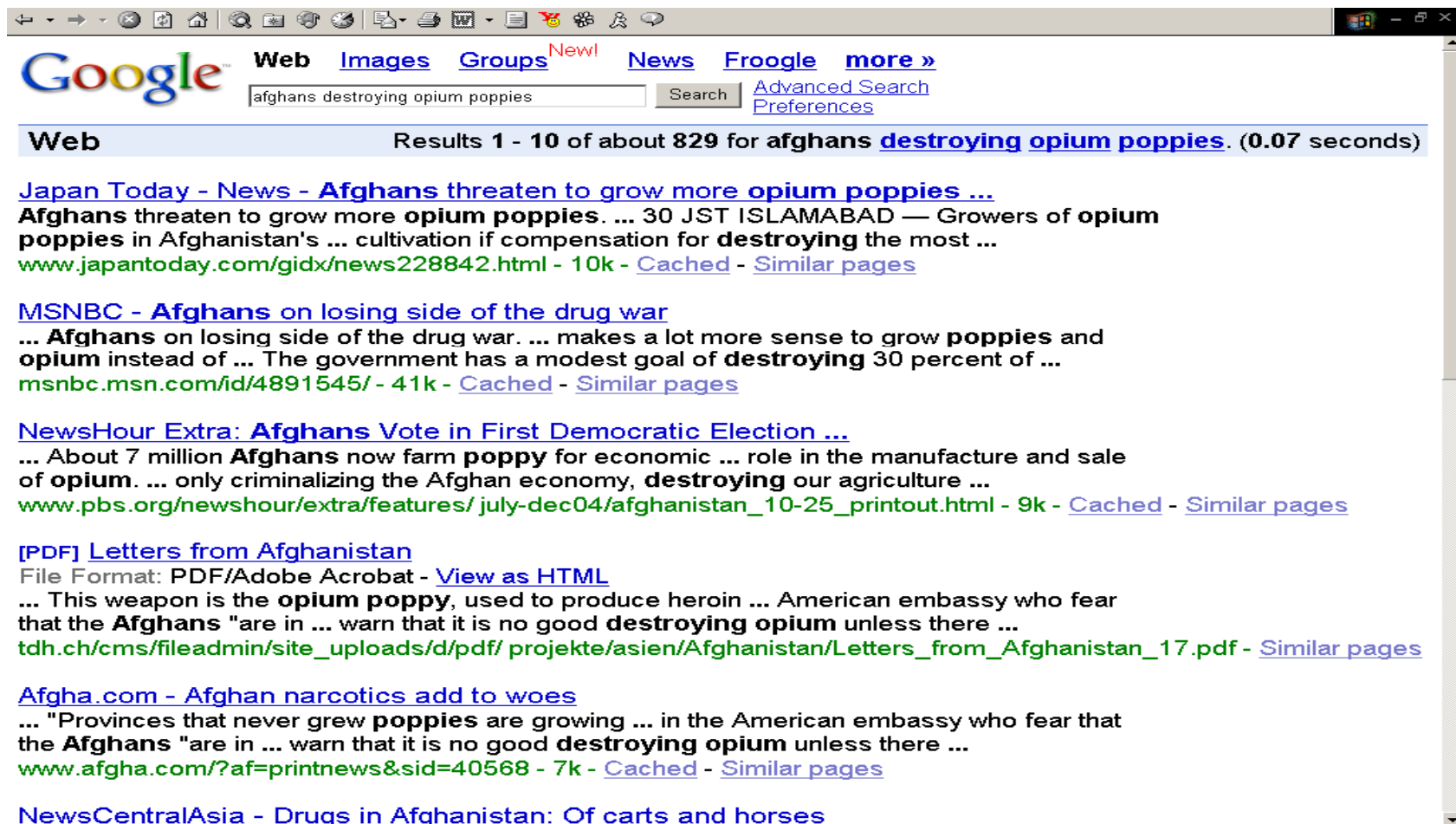
- [Native American Images - American Indian North America Tribe Map](#)
- Native American Images American **Indian** North **America** **Tribe** Map Click here to view more images ... History Hotline | Iraqi War | Korean War | **Latin** Americans ...
 - www.nativeamericans.com/NativeAmericanImages6.htm
 - [Cached page](#)
- [Resources for](#)
- ... 152t.), Panama (126t.), Paraguay (67t.), Surinam (10t.), and Venezuela (331t.) (t.=thousand). - **Indian Tribes** in **Latin America**
 - www.xmission.com/~amauta/tribes.htm
 - [Cached page](#)
- [Latin America Community Assistance Foundation - LACA](#)
- The Tarahumara Indians are the most primitive of all **Indian tribes** in North **America**, and are the least touched by modern society.
 - www.lacafoundation.org/?page_id=58
 - [Cached page](#)
- [Latin America Tour Set for Curtis Photos of North America Tribes](#)
- 28 September 2005. **Latin America** Tour Set for Curtis Photos of North **America** **Tribes**. Famed photographer recorded **Indian** tribal life in 19th, early 20th century
 - www.america.gov/st/washfile-english/2005/September/20050928134700GLnesnoMO.2225763.html
- [Latin America // Current](#)
- Current TV **Latin America** category, discover popular **Latin America** stories, news and ... of the Amazon jungle, a land conflict between rice farmers and a handful of **Indian tribes** ...
 - current.com/topics/75844112_latin_america
 - [Cached page](#)
- [Bloomberg.com: Latin America](#)
- May 30 (Bloomberg) -- Brazil's National **Indian** Foundation has discovered an **Indian tribe** in the Amazon that hasn't had contact with civilization in a rare sighting of the few ...
 - www.bloomberg.com/apps/news?pid=20601086&sid=aSrj5wfHW.CQ&refer=latin_america

Personalized focused search (wikipedia.cognition)

- **Indian Latin-America tribe:** 249 files —
- **William Curtis Farabee**
- The volumes that Farabee produced from his travels include Indian Tribes of Eastern Peru based on his first trip in 1906-1908 (Obituary, 1925).
- [Direct link \(no highlighting\)](#)
- **Mexican Texas**
- Settlers were empowered to create their own militias to help control hostile Indian tribes. Texas faced raids from both the Apache and Comanche tribes, [...]
- [Direct link \(no highlighting\)](#)
- **Temecula, California**
- The Luiseño and Cahuilla tribes were involved, rather bloodily, in the local battles of the Mexican-American War during the following years.
- [Direct link \(no highlighting\)](#)
- **Kaweah Indian Nation**
- Recently, scam artists have sold purported citizenships in the non-recognized tribe, particularly to Mexican nationals who have entered the US illegally.¹ [...]
- [Direct link \(no highlighting\)](#)
- **Flag of Puerto Rico**
- The tribal nation flag of the Jatibonicu Taino Indians of Borikén, represents the Jatibonicu Taino tribe's original pre-Columbian territories of [...]
- [Direct link \(no highlighting\)](#)
- **Maina Indians**
- The Maina Indians are a group of tribes constituting a distinct linguistic stock, the [...] along the north bank of the Marañón River in South America
- [Direct link \(no highlighting\)](#)
- **Erie (tribe)**
- ^ Ebooks by Google: "Handbook of American Indians North of Mexico" By Frederick Webb Hodge <http://books.google.com/books?>
- [Direct link \(no highlighting\)](#)
- **Miccosukee**
- [1] Other members went on to form the Miccosukee Tribe of Indians of Florida, which was not recognized by Fidel Castro's Cuban government in 1959. The [...]
- [Direct link \(no highlighting\)](#)
- **New Tribes Mission**
- In Paraguay in 1979 and 1986, New Tribes Mission was accused of assisting in the forcible contact of nomadic Ayoreo Indians.
- [Direct link \(no highlighting\)](#)

Example: Semantically precise search for relations/events

Query: *afghans destroying opium poppies*



The screenshot shows a Google search interface with the query "afghans destroying opium poppies" entered in the search box. The search results are displayed under the "Web" tab, showing the first 10 results out of approximately 829. The search took 0.07 seconds. The results include news articles from Japan Today, MSNBC, NewsHour Extra, and Afgha.com, as well as a PDF document from tdh.ch. The search results are semantically precise, focusing on the relationship between "afghans" and "destroying opium poppies".

Web Results 1 - 10 of about 829 for **afghans destroying opium poppies**. (0.07 seconds)

[Japan Today - News - Afghans threaten to grow more opium poppies ...](#)
Afghans threaten to grow more **opium poppies**. ... 30 JST ISLAMABAD — Growers of **opium poppies** in Afghanistan's ... cultivation if compensation for **destroying** the most ...
www.japantoday.com/gidx/news228842.html - 10k - [Cached](#) - [Similar pages](#)

[MSNBC - Afghans on losing side of the drug war](#)
... **Afghans** on losing side of the drug war. ... makes a lot more sense to grow **poppies** and **opium** instead of ... The government has a modest goal of **destroying** 30 percent of ...
msnbc.msn.com/id/4891545/ - 41k - [Cached](#) - [Similar pages](#)

[NewsHour Extra: Afghans Vote in First Democratic Election ...](#)
... About 7 million **Afghans** now farm **poppy** for economic ... role in the manufacture and sale of **opium**. ... only criminalizing the Afghan economy, **destroying** our agriculture ...
www.pbs.org/newshour/extra/features/july-dec04/afghanistan_10-25_printout.html - 9k - [Cached](#) - [Similar pages](#)

[\[PDF\] Letters from Afghanistan](#)
File Format: PDF/Adobe Acrobat - [View as HTML](#)
... This weapon is the **opium poppy**, used to produce heroin ... American embassy who fear that the **Afghans** "are in ... warn that it is no good **destroying opium** unless there ...
tdh.ch/cms/fileadmin/site_uploads/d/pdf/projekte/asien/Afghanistan/Letters_from_Afghanistan_17.pdf - [Similar pages](#)

[Afgha.com - Afghan narcotics add to woes](#)
... "Provinces that never grew **poppies** are growing ... in the American embassy who fear that the **Afghans** "are in ... warn that it is no good **destroying opium** unless there ...
www.afgha.com/?af=printnews&sid=40568 - 7k - [Cached](#) - [Similar pages](#)

[NewsCentralAsia - Drugs in Afghanistan: Of carts and horses](#)

Needs The robust textual inference task

[Dagan, Glickman, and Magnini 2005]

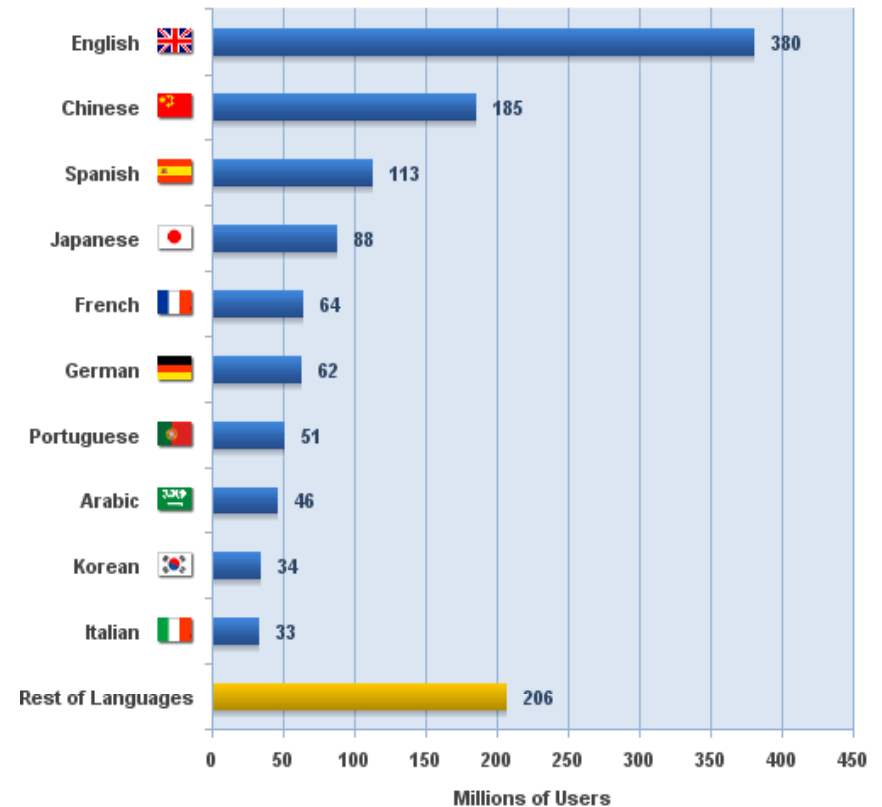
- Task: Does text T justify an inference to hypothesis H ?
 - On the assumption that some piece of text (T) is true, does this imply the truth of some other hypothesis text (H)?
 - *Sydney was the host city of the 2000 Olympics* →
 - *The Olympics have been held in Sydney* TRUE
- In practice:
 - An informal, intuitive notion of inference
 - Not strict logic; incorporates pragmatics, and default assumptions
 - Focus on local inference steps, not long chains of deduction
 - Includes basic world knowledge but not highly technical material
 - Relevance logic: text must inform hypothesis

India Wide Cross Lingual Information Access (CLIA) Endeavour

Motivation

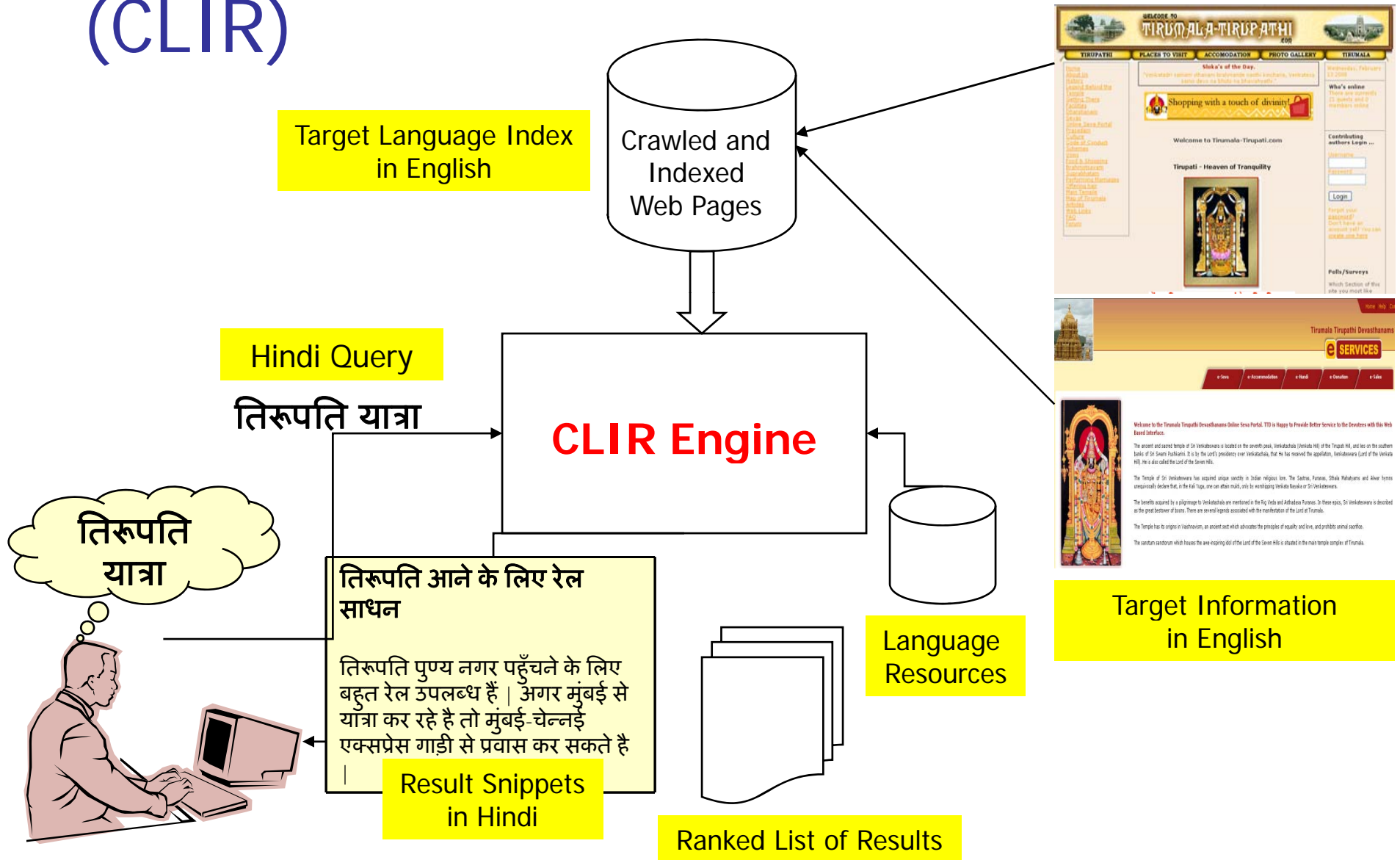
- English still the most dominant language on the web
 - ❖ Contributes 72% of the content
- Number of non-English users steadily rising all over the world
- English penetration in India
 - ❖ Estimated to be around 3-4%
 - ❖ Mostly the urban educated class
- Need to enable access to above information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

Top 10 Internet Languages - November 2007



Source: www.internetworldstats.com
Copyright © 2008, Miniwatts Marketing Group

Cross Language Information Retrieval (CLIR)



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

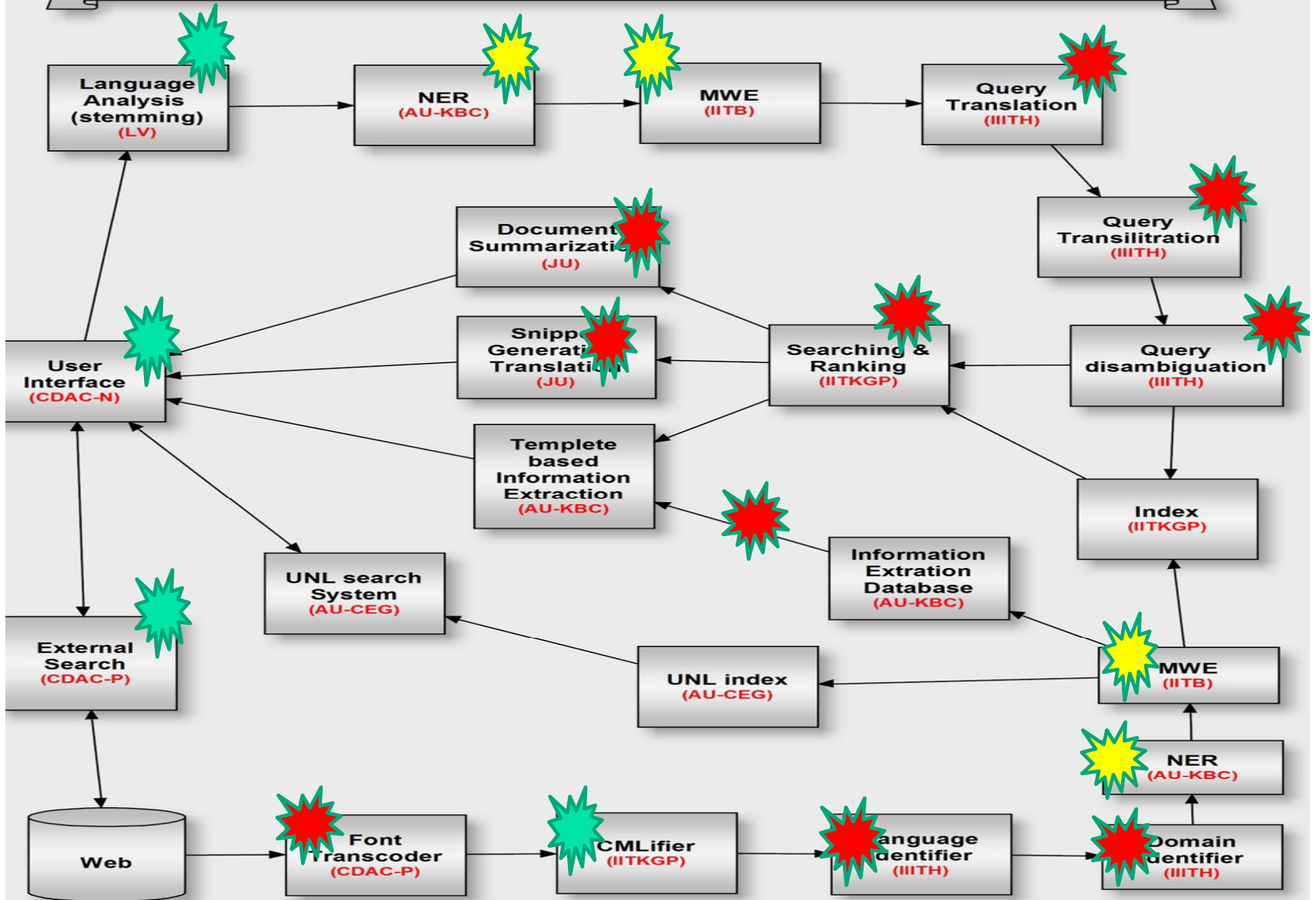
Cross Language Information Access (CLIA) Consortia Project

- 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

Portal

- **Public portal released** at <http://www.clia.iitb.ac.in/clia-beta-ext/> in September 2009. (Outside IITB)
- **Public portal released** at <http://www.clia.iitb.ac.in:8080/clia-beta-ext/> in September 2009. (Inside IITB)

Cross Lingual Information Access



Recent Press Coverage

13/2/09

The Indian EXPRESS ****

www.expressindia.com

MAHARASHTRA

At last, Net may respond to India's multilingual needs

■ Enter your query in Marathi; get snippets of all retrieved documents in English, Hindi & Marathi

UPNEET PANSARE

MUMBAI, FEBRUARY 12

A CONSORTIUM of 11 academic and research institutions and industry partners across India including the Indian Institute of Technology, Bombay (IITB), are currently finetuning a project that will enable users retrieve information from websites in their preferred Indian language (Bengali, Hindi, Marathi, Punjabi, Tamil and Telugu).

The project, the Cross Lingual Information Access Project (CLIAP) funded by the Union Communications and Information Technology Ministry, Department of Infor-

mation Technology, will be available for public use on the website of the ministry or may have a separate URL of its own around May.

Resource creation for the project has been divided among institutes of the likes of IITB, IIT Kharagpur, Indian Institute of Information Technology (IIIT) Hyderabad, Anna University Chennai, Indian Statistical Institute (ISI) Kolkata, Jadavpur University at Kolkata, Centre for Development of Advancement Computing Pune, C-DAC Noida, Utkal University Bhubaneswar and Delhi Institute of Technology New Delhi. Each of these institutes is working on the local resource creation of the local language there (for instance, IITB is responsible for resource

creation in Marathi, Hindi and English).

Simply put, a user will be able to enter a query in one Indian language and access documents available in the language of the query and English. Results of the query will be presented to the user in the language of the query. The results can also be presented in the language in which the information originally resided. So, if the source language selected is Marathi, the user can enter the query in Marathi. The CLIA system searches for documents in English, Hindi and Bengali either from the web or the local site and snippets of the retrieved documents are displayed in English/Hindi and Marathi. The user can choose from different domains

though the project is initially restricted to tourism and health domains. Since India is a multilingual country the need for such a system becomes more evident. Pushpak Bhattacharyya, professor, department of Computer Science and Engineering (CSE) said, "One of the most important aim of the project would be to introduce the agricultural domain and extend these services to the farming community in the country."

According to Vishal Vachani, a CSE M Tech student currently working on this project, the project gives snippet information of the search results in the preferred language. Work, meanwhile is on to provide fully translated webpages in the future.

Hindustan Times

The Cross Language Information Access query screen in Marathi (below) and the answers (right)



Multi-lingual search

PROFESSOR PUSHPAK BHATTACHARYA,
Computer Science Department

What

The Cross Language Information Access (CLIA) program.

What's cool

Get information on popular Indian tourist places in the language you speak. If you ask a question in Marathi, the program returns responses in the same language.

Prof says

CLIA is important in a multi-lingual country like ours where only five per cent of the population knows English. The language barrier needs to be broken to serve the information needs of all Indians.

Besides English, the program works with Hindi, Marathi, Bengali, Telegu, Tamil and Punjabi.

It is currently available in the domain of tourism but will soon include agriculture and health as well.

IR Basics

(mainly from *R. Baeza-Yates and B. Ribeiro-Neto. Modern Information Retrieval* Addison-Wesley, Wokingham, UK, 1999.

and

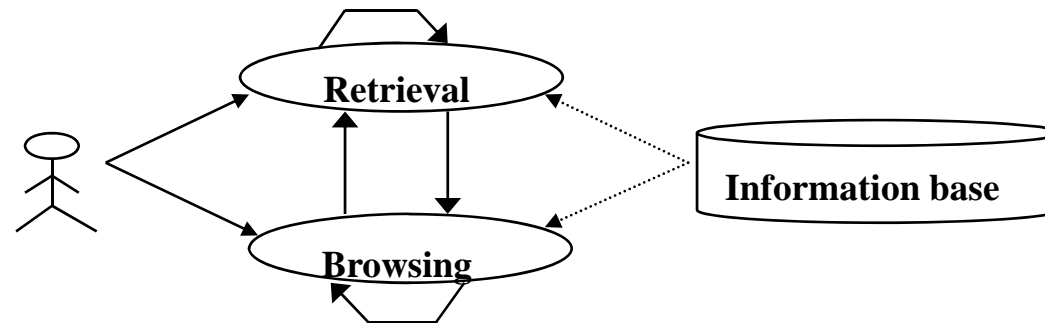
Christopher D. Manning, Prabhakar Raghavan and Hinrich Schütze, *Introduction to Information Retrieval*, Cambridge University Press. 2008.)

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!

Basic Concepts

- The User Task



- Retrieval

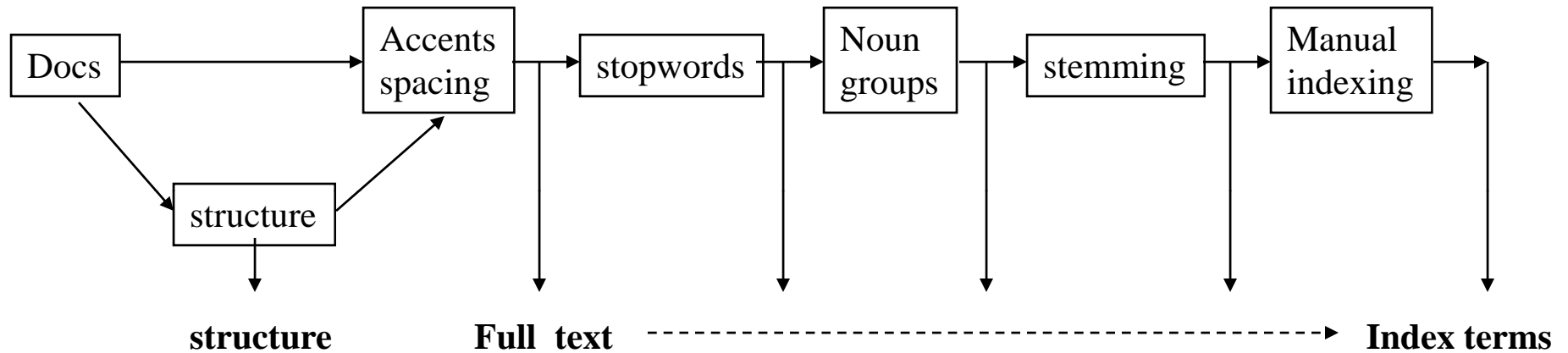
- information or data
- purposeful

- Browsing

- glancing around
- F1; cars, Le Mans, France, tourism

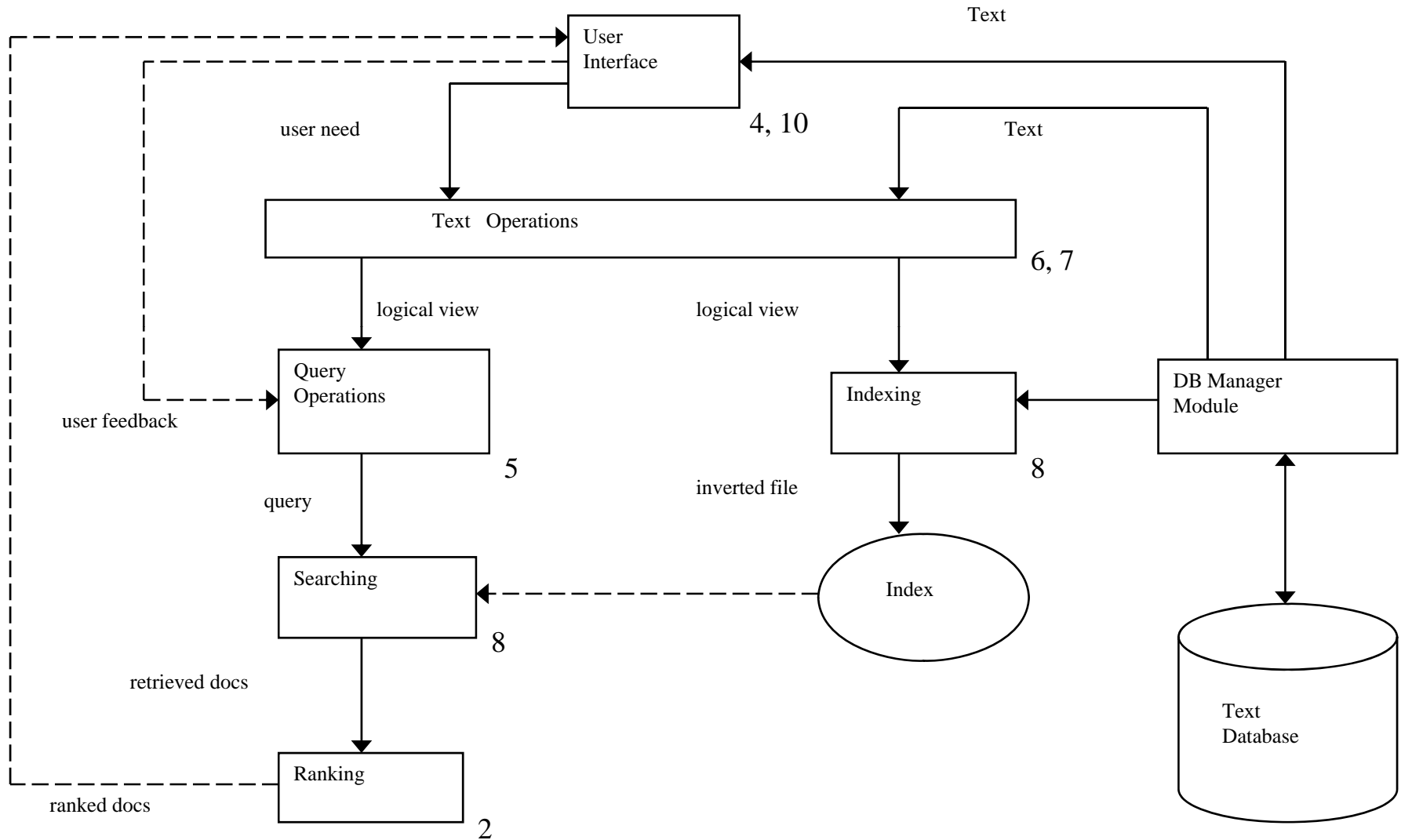
Basic Concepts

- Logical view of the documents



- Structure and meaning of the document lost in the representation

The Retrieval Process



IR Models

