India's Cross Lingual Information Access

Presented by CLIA consortium of India

2006 Phase I started (December) → 2007 Basic India Search designed → 2008 Language Analysis Enhancements → 2009 India Search Release → 2010 Deployments of Sandhaan → 2011 Phase II started . . . (March)

Introduction

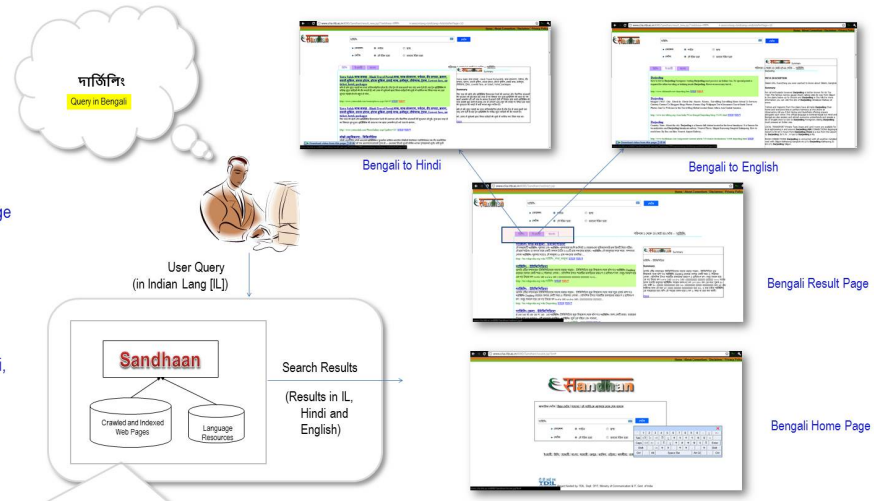
Scope:

- A mission mode Project executed by a consortium of academic and research institutions and industry partners.
- Funded by Ministry of Communication and Information Technology, Government of India.
- Aiming to develop a Search Engine with the support of 9 Indian languages.

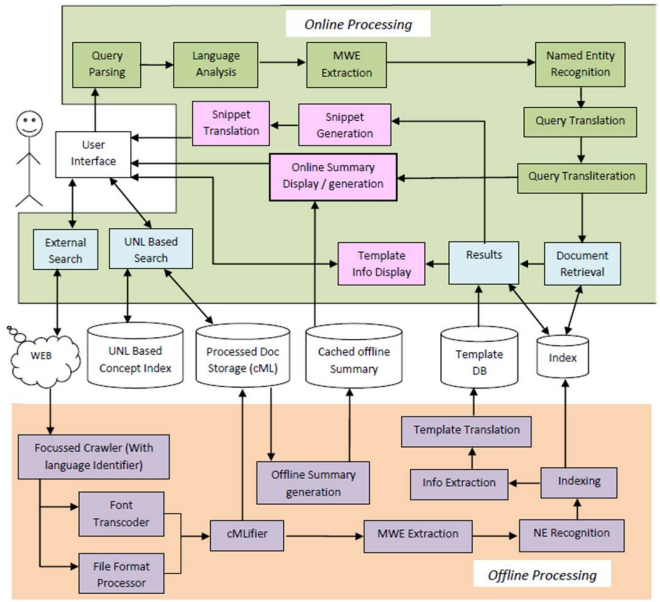
Working Principle:

1. A user will be able to give a query in one Indian language and
2. User is able to access documents available in
   a. the language of the query,
   b. Hindi (if the query language is not Hindi), and
   c. English
3. **Query Languages: 9**
   - ❖ Assamese, Bangla, English, Gujarati, Hindi, Marathi, Oriya, Punjabi, Tamil, Telugu
   - ❖ **Results in**
     Query Language + Hindi + English
   - ❖ **Domain**
     Tourism and Health

Working Of CLIA Engine



CLIA Architecture



Key Achievements

- Fully operational Cross-lingual web-based search system
- Development of valuable language resources (data and tools) in all the involved languages
- Development of data for Indian Language CLIR system evaluation (FIRE)
- All resources made accessible by Department of Information Technology (DIT) through the data center
- Forum for Information Retrieval Evaluation (FIRE) – workshop series for CLIR evaluation for Indian Languages.
- Development of a strong and connected research community around Cross Language Information Retrieval (CLIR) in Indian languages
- The CLIA system can be accessed at the following link: http://www.clia.iitb.ac.in/sandhan-1.0
- Research Publications: 33 in top level conferences
- Students graduated working on CLIA: 5 PhDs, 26 masters
- Special issue of Vishwabhrat coming up
- Tools and resources copyrighted

Resources Developed

- ❖ Stemmer for all the 9 Indian Languages
- ❖ Font Transcoder
- ❖ Translation and Transliteration
- ❖ Summary and Snippet Generation and Translation
- ❖ Query Processing and Query Translation
- ❖ Multiword Expression identification
- ❖ Named Entity Identification

Future Outlook

- ❖ Research (many problems and ideas being worked on)
- ❖ Public release
  - Search for industry partner
  - MLIR for 6 languages by Jan 2012: roadmap
  - Role of Ministry (Data center, evangelization, Tour operators, public portal)
  - Technology (bandwidth, Server with Spec)
  - Stress testing
  - Revenue model

FIRE

- ❖ The Forum for Information Retrieval Evaluation (FIRE) provides a common evaluation infrastructure for comparing the performance of Sandhaan with commercial search engines like Google, Yahoo, Rediff.
- ❖ FIRE investigates evaluation methods for Information Access techniques and methods for constructing a reusable large-scale data set for ILIR experiments.
- ❖ The FIRE evaluation methods are similar to TREC standards but focused on retrieval in Indian languages.

Consortium Members

IIT Bombay | IIT Kharagpur | Anna University (AU-KBC | AU-CEG) | CDAC Noida | CDAC Pune | DAIICT Gandhinagar | IIIT Hyderabad | ISI Kolkata | Jadavpur University | IIIT Bhubaneswar | Guwahati University