Analysis of Cross Language Information Retrieval methods

Introduction to Cross Language Information Retrieval (CLIR)

- CLIR is a subfield of information retrieval dealing with retrieving information written in a language different from the language of the user's query.
- Information Retrieval systems should be capable of searching for information in multiple languages
- Cross Language Information Retrieval (CLIR) is an intersection of Machine Translation and Information Retrieval

Motivation

- The need to acquire information even if it's not available in the user's native language
- CLIR may bridge the gap between the desire to obtain information and unavailability or under-availability of such information in their native language.
- Retrieve information from a multilingual collection using a query in a single language
- Locate documents in a multilingual collection of scanned pages

Importance of CLIR

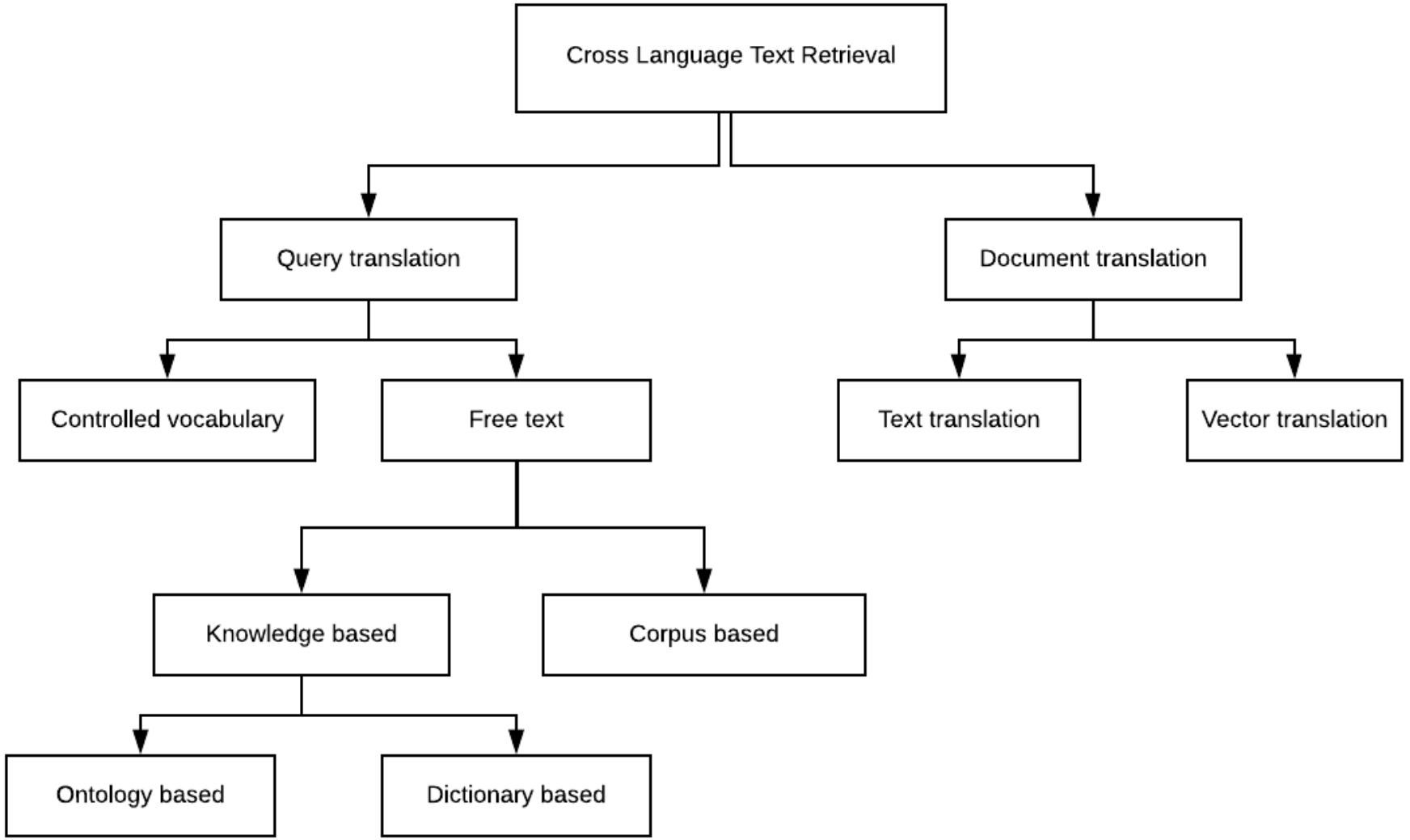
- CLIR research is important for global information exchange and sharing of knowledge
  - National Security
  - Foreign Patent information access
  - Medical information access for patients
  - Sentiment analysis
  - Information Extraction

Issues of CLIR

- How to convert a term to another language?
- Which of the possible translations should be retained?
- How to properly weigh the importance of translation alternatives?

Design decisions

- What to index?
  - Free text or controlled vocabulary
- What to translate?
  - Queries or documents
- Where to get translation knowledge?
  - Dictionary, ontology, training corpus


Cross Language Text Retrieval
Query translation
Document translation
Controlled vocabulary
Free text
Text translation
Vector translation
Knowledge based
Corpus based
Ontology based
Dictionary based

Query VS Document translation

- Query translation
  - Very efficient for short queries
    - Not as big an advantage for relevance feedback
  - Hard to resolve ambiguous query terms

- Document translation
  - Slow, but only need to do it once per document
    - Poor scale-up to large number of languages

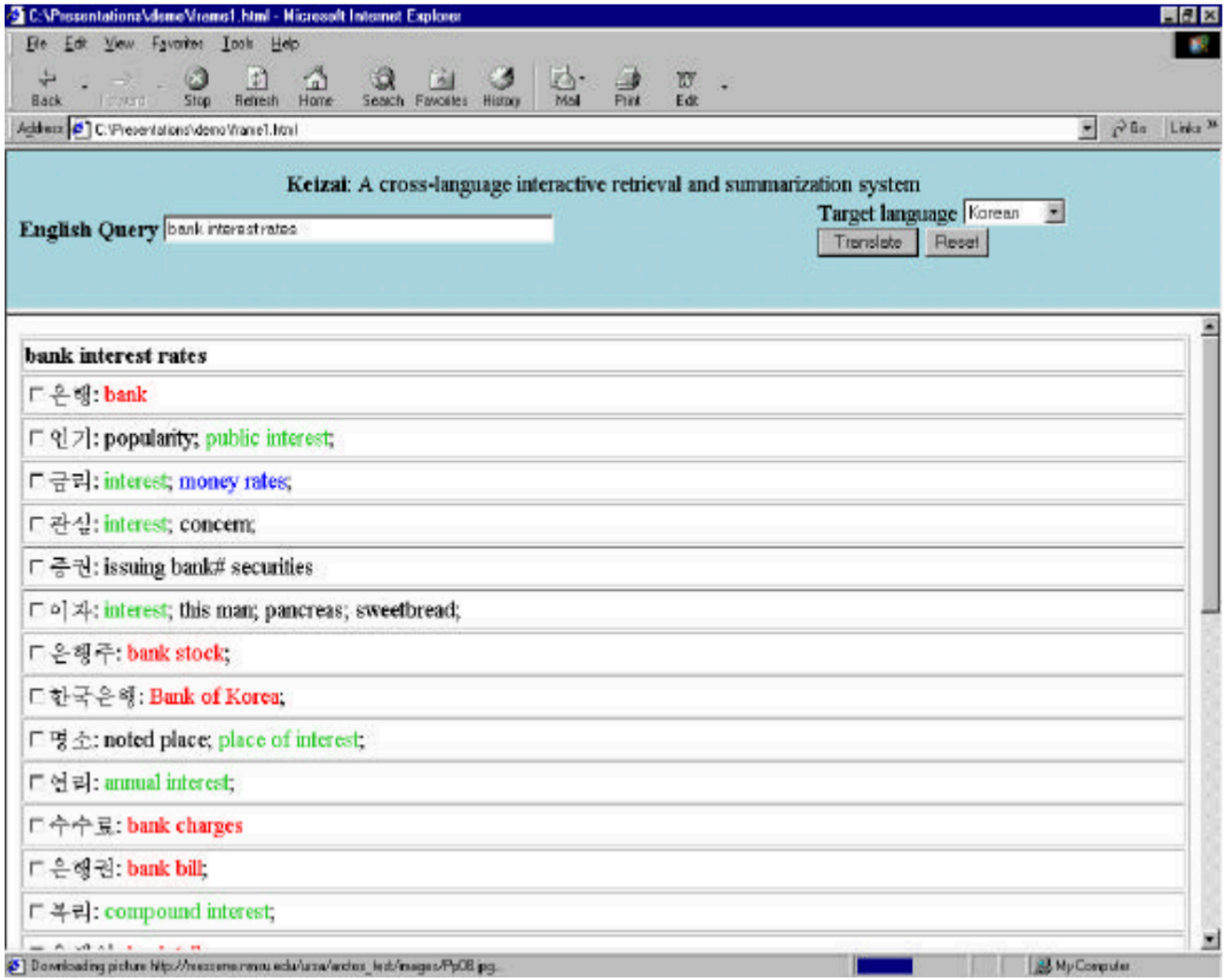
Recent trends in CLIR research

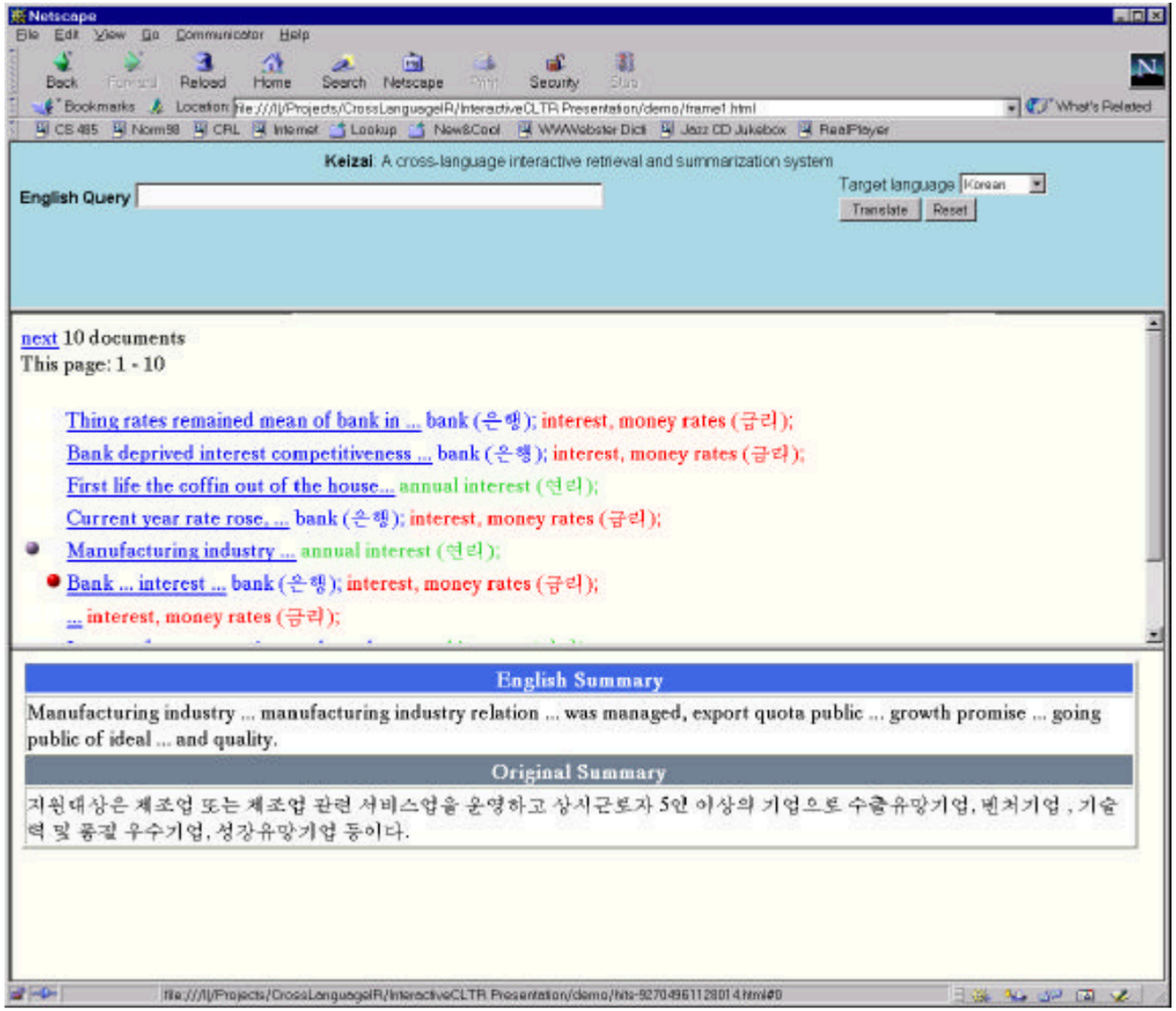
- Keizai CLTR system
- English – Hindi CLIR system
- Cross Lingual Information Retrieval and Delivery using community mobile networks
- Ontologies

Keizai CLTR system

- Uses the query translation approach
- User inputs English query, system searches Japanese and Korean web data
- Displays English summaries on top ranking documents
- User needs to accurately judge which foreign language documents are relevant to their query
- Provides extended English definitions of query terms alongside Japanese or Korean translations

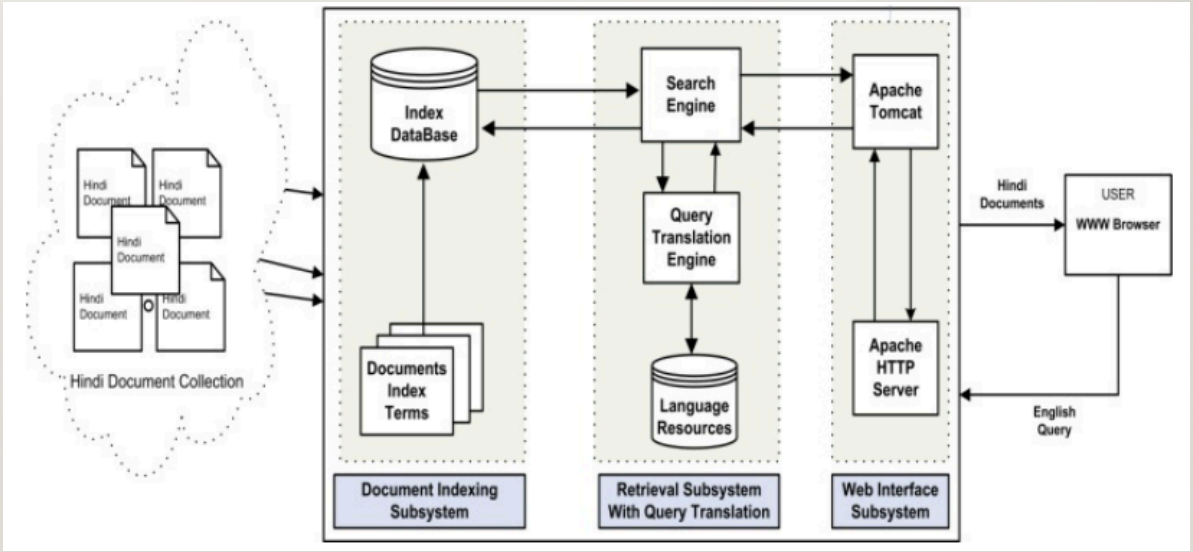
KEIZAI QUERY TERM SELECTION




Netscape
File Edit View Go Communicator Help
Back Forward Reload Home Search Netscape Print Security Stop
Bookmarks Location: file:///I|/Projects/CrossLanguageIR/InteractiveCLTR Presentation/demo/frame1.html What's Related
CS 485 Norm98 CRL Internet Lookup New&Cool WWWebster Dict Jazz CD Jukebox RealPlayer
Keizai: A cross-language interactive retrieval and summarization system
English Query
Target language: Korean
Translate Reset
next 10 documents
This page: 1 - 10
Thing rates remained mean of bank in ... bank (은행); interest, money rates (금리);
Bank deprived interest competitiveness ... bank (은행); interest, money rates (금리);
First life the coffin out of the house... annual interest (연리);
Current year rate rose, ... bank (은행); interest, money rates (금리);
Manufacturing industry ... annual interest (연리);
Bank ... interest ... bank (은행); interest, money rates (금리);
... interest, money rates (금리);
English Summary
Manufacturing industry ... manufacturing industry relation ... was managed, export quota public ... growth promise ... going public of ideal ... and quality.
Original Summary
지원대상은 제조업 또는 제조업 관련 서비스업을 운영하고 상시근로자 5인 이상의 기업으로 수출유망기업, 벤처기업 , 기술력 및 품질 우수기업, 성장유망기업 등이다.
file:///I|/Projects/CrossLanguageIR/InteractiveCLTR Presentation/demo/hits-927049611280l4.html#0

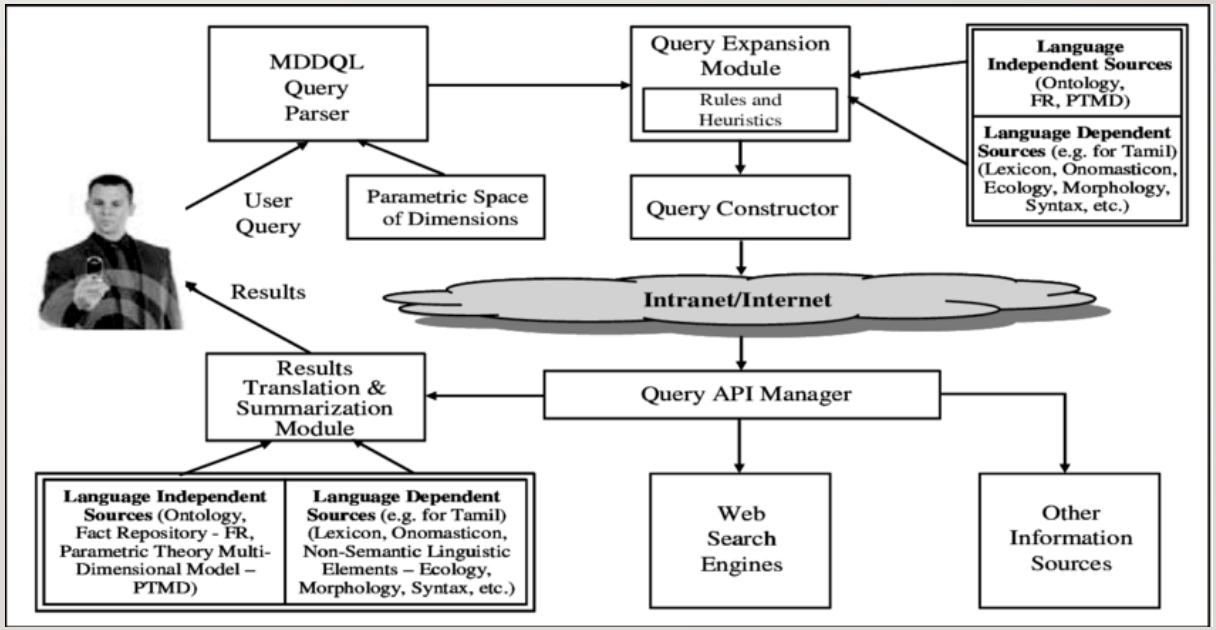
English-Hindi CLIR system

- CLIR system developed using Managing Gigabytes (MG) retrieval system as the base IR system
- Converts query in English to Hindi
- Publicly available online bilingual dictionary 'Shabdanjali' used for query translation
- Quality of translation depends on the quality of dictionary



Cross lingual information retrieval and delivery using community mobile networks

- Searches appropriate content and summarizes using a content-specification meta language
- Focuses on querying the Web in languages other than English, namely south Indian languages including Tamil.
- Retrieves relevant documents, translate, summarize and present the information to user in Tamil language



Ontologies

- Ontology is a formal, explicit specification of a shared conceptualization.
- Retrieving English documents relevant to Persian queries using Bilingual ontology to annotate the documents and queries
- A bilingual ontology consists of ontology and a bilingual dictionary
- Ontology is used to expand the query with related terms in pre and post translation expansion and the combined approach significantly improves cross-lingual performance

Ontologies

- Researchers analyzed query translation in cross lingual IR based on feature vectors and usage of context information
- Using information external to the query, such as the ontologies, the effect of disambiguation can be reduced.

Future scope of CLIR systems

- Availability for all languages
  - CLIR available only for top commonly used languages
  - Other languages are left out

- Multi-lingual IR
  - This type of IR will not be restricted only to two languages
  - Will include multiple languages to broaden the search results

References

[1] Ogden, William & Cowie, James & Davis, Mark & Ludovik, Eugene & Nirenburg, Sergei & Sharples, Nigel. (2000). Keizai: An Interactive Cross-Language Text Retrieval System.
[2] Raghunathan, Shriram & Sugumaran, Vijayan & Kapetanios, Epaminondas. (2007). Cross-Lingual Information Retrieval and Delivery Using Community Mobile Networks. 320 - 325. 10.1109/ICDIM.2007.369217.
[3] A. Seetha, S. Das and M. Kumar, "Evaluation of the English-Hindi Cross Language Information Retrieval System Based on Dictionary Based Query Translation Method," 10th International Conference on Information Technology (ICIT 2007), Orissa, 2007, pp. 56-61.
[4] V. Pemawat, A. Saund and A. Agrawal, "Hindi - English based cross language Information Retrieval system for Allahabad Museum," 2010 International Conference on Signal and Image Processing, Chennai, 2010, pp. 153-157.
[5] B. A. Kumar, "Profound Survey on Cross Language Information Retrieval Methods (CLIR)," 2012 Second International Conference on Advanced Computing & Communication Technologies, Rohtak, Haryana, 2012, pp. 64-68.
[6] Jian-Yun Nie, "Cross-Language Information Retrieval," in Cross-Language Information Retrieval , Morgan & Claypool, 2010
[7] P. Liu, Z. Zheng and Q. Su, "Cross-Language Information Retrieval Based on Multiple Information," 2018 IEEE/WIC/ACM International Conference on Web Intelligence (WI), Santiago, 2018, pp. 623-626.