

Winter School on Information Retrieval Systems and Experimentation

Mandar Mitra

Indian Statistical Institute

- What this school is about
- Experimentation, evaluation and systems
- What is FIRE and why you should care

Assumptions:

- You are interested in Information Retrieval (IR).
- You have ideas about how to improve quality of search results.
- You are afraid you will have to write a complete IR system.
- You are not sure where to get test data.

Assumptions:

- You are interested in Information Retrieval (IR).
- You have ideas about how to improve quality of search results.
- You are afraid you will have to write a complete IR system.
- You are not sure where to get test data.

How we hope to help:

- Introduce available IR systems and available datasets
- Experimental methodology and evaluation metrics

Limitations:

- Restricted to the basic retrieval task
 - no advanced tasks, e.g. question-answering, summarization, etc.
- Restricted to static corpora
 - no Web search

The Cranfield method (CLEVERDON ET AL., 60S)

- Document collection
- Query / topic collection
- Relevance judgments - information about which document is relevant to which query
 - most tedious component

The Cranfield method (CLEVERDON ET AL., 60S)

- Document collection
- Query / topic collection
- Relevance judgments - information about which document is relevant to which query
 - most tedious component

Evaluation forums

- TREC
- CLEF
- NTCIR
- FIRE

`http://trec.nist.gov`

- Organized by NIST every year since 1992
- Typical tasks
 - adhoc
 - user enters a search topic for a one-time information need
 - document collection is static
 - routing/filtering
 - user's information need is persistent
 - document collection is a stream of incoming documents
 - question answering

■ Documents

■ Genres:

- news (AP, LA Times, WSJ, SJMN, Financial Times, FBIS)
- govt. documents (Federal Register, Congressional Records)
- technical articles (Ziff Davis, DOE abstracts)

- Size: 0.8 million documents – 1.7 million web pages
(cf. Google indexes several billion pages)

■ Topics

- title
- description
- narrative

<http://www.clef-campaign.org/>

- CLIR track at TREC-6 (1997), CLEF started in 2000
- Objectives:
 - to provide an infrastructure for the testing and evaluation of information retrieval systems operating on European languages in both monolingual and cross-language contexts
 - to construct test-suites of reusable data that can be employed by system developers for benchmarking purposes
 - to create an R&D community in the cross-language information retrieval (CLIR) sector

- Monolingual retrieval
- Bilingual retrieval
 - queries in language X
 - document collection in language Y
- Multi-lingual retrieval
 - queries in language X
 - multilingual collection of documents (e.g. English, French, German, Italian)
 - results include documents from various collections and languages in a single list
- Other tasks: spoken document retrieval, image retrieval

<http://research.nii.ac.jp/ntcir/index-en.html>

- Started in late 1997
- Held every 1.5 years at NII, Japan
- Focus on East Asian languages (Chinese, Japanese, Korean)
- Tasks
 - cross-lingual retrieval
 - patent retrieval
 - geographic IR
 - opinion analysis

<http://www.irsi.res.in>

- Forum for Information Retrieval Evaluation
- Evaluation component of a DIT-sponsored, consortium mode project
- Assigned task: create a portal where
 - a user will be able to give a query in one Indian language; s/he will be able to access documents available in the language of the query, Hindi (if the query language is not Hindi), and English,
 - all presented to the user in the language of the query.
- Languages: Bangla, Hindi, Marathi, Punjabi, Tamil, Telugu

- To encourage research in South Asian language Information Access technologies by providing reusable large-scale test collections for ILIR experiments
- To provide a common evaluation infrastructure for comparing the performance of different IR systems
- To explore new Information Retrieval / Access tasks that arise as our information needs evolve, and new needs emerge
- To investigate evaluation methods for Information Access techniques and methods for constructing a reusable large-scale data set for ILIR experiments.
- To build language resources for IR and related language processing tasks

- To encourage research in South Asian language Information Access technologies by providing reusable large-scale test collections for ILIR experiments
- To provide a common evaluation infrastructure for comparing the performance of different IR systems
- To explore new Information Retrieval / Access tasks that arise as our information needs evolve, and new needs emerge
- To investigate evaluation methods for Information Access techniques and methods for constructing a reusable large-scale data set for ILIR experiments.
- To build language resources for IR and related language processing tasks

Build a strong, friendly community!

- Ad-hoc monolingual retrieval (repeat)
 - Bengali, Hindi Marathi and English
- Ad-hoc cross-lingual document retrieval (repeat)
 - documents in Bengali, Hindi, Marathi, and English
 - queries in Bengali, Hindi, Marathi, Tamil, Telugu , Gujarati and English
 - Roman transliterations of Bengali and Hindi topics
- Retrieval and classification from mailing lists and forums (new)
 - pilot task being offered by IBM India Research Lab
- Ad-hoc Wikipedia-entity retrieval from news documents (new)
 - pilot task being offered by Yahoo! Labs, Bangalore.

Documents

- Bengali: Anandabazar Patrika (123,047 docs)
- Hindi: Dainik Jagran (95,215 docs) +
Amar Ujala (54,266 docs)
- Marathi: Maharashtra Times, Sakal (99,275 docs)
- English: Telegraph (125,586 docs)

- All from the Sep 2004 - Sep 2007 period
- All content converted to UTF-8
- Minimal markup

Topics

- Topics: mix of national and international issues
- Queries formulated parallelly in Bengali, Hindi by browsing the corpus
- Refined based on initial retrieval results
 - ensure minimum number of relevant documents per query
 - balance easy, medium and hard queries
- Translated into Marathi, Tamil, Telugu , Gujarati and English
- TREC format (title + desc + narr)

Example:

<title> Nobel theft

<desc>

Rabindranath Tagore's Nobel Prize medal was stolen from Santiniketan. The document should contain information about this theft.

<narr>

A relevant document should contain information regarding the missing Nobel Prize Medal that was stolen along with some other artefacts and paintings on 25th March, 2004. Documents containing reports related to investigations by government agencies like CBI / CID are also relevant, as are articles that describe public reaction and expressions of outrage by various political parties.

Participants

Institute	Country	# runs submitted
AU-KBC	India	2
Dublin City U.	Ireland	17
IBM	India	2
IIT Bombay (1)	India	30
IIT Bombay (2)	India	3
Jadavpur U.	India	2
MANIT	India	9
Microsoft Research	India	32
U. Neuchatel	Switzerland	18
U. North Texas	USA	8
U. Tampere	Finland	6
11 (9 @ FIRE 2008)	TOTAL	129 (up from 64 @ FIRE 2008)

- Retrieval and classification from mailing lists and forums
 - pilot task being offered by IBM India Research Lab
- Ad-hoc Wikipedia-entity retrieval from news documents
 - pilot task being offered by Yahoo! Labs, Bangalore.

Hope to see you at FIRE 2011!

`http://isical.ac.in/~fire`

`http://www.irsi.res.in`