CS344: Introduction to Artificial Intelligence

Vishal Vachhani

M.Tech, CSE

Lecture 34-35: CLIR and Ranking in IR

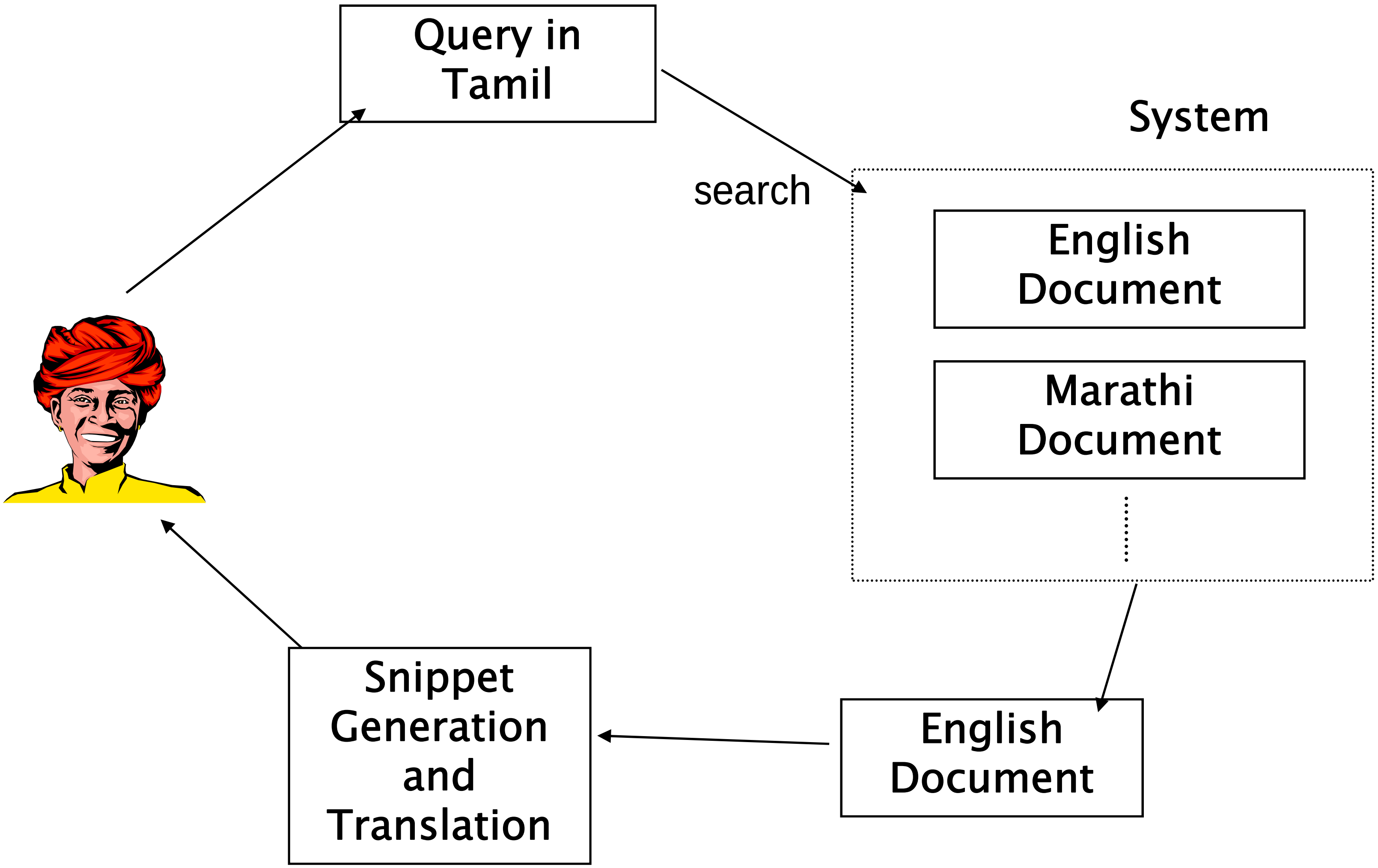
Road Map

- Cross Lingual IR
  - Motivation
  - CLIA architecture
  - CLIA demo
- Ranking
  - Various Ranking methods
  - Nutch/lucene Ranking
  - Learning a ranking function
  - Experiments and results

Cross Lingual IR

- Motivation
  - Information unavailability in some languages
  - Language barrier
- Definition:
  - **Cross-language information retrieval (CLIR)** is a subfield of information retrieval dealing with retrieving information written in a language different from the language of the user's query (wikipedia)
- Example:
  - A user may ask query in Hindi but retrieve relevant documents written in English.

Why CLIR?



Query in Tamil

search

System

English Document

Marathi Document

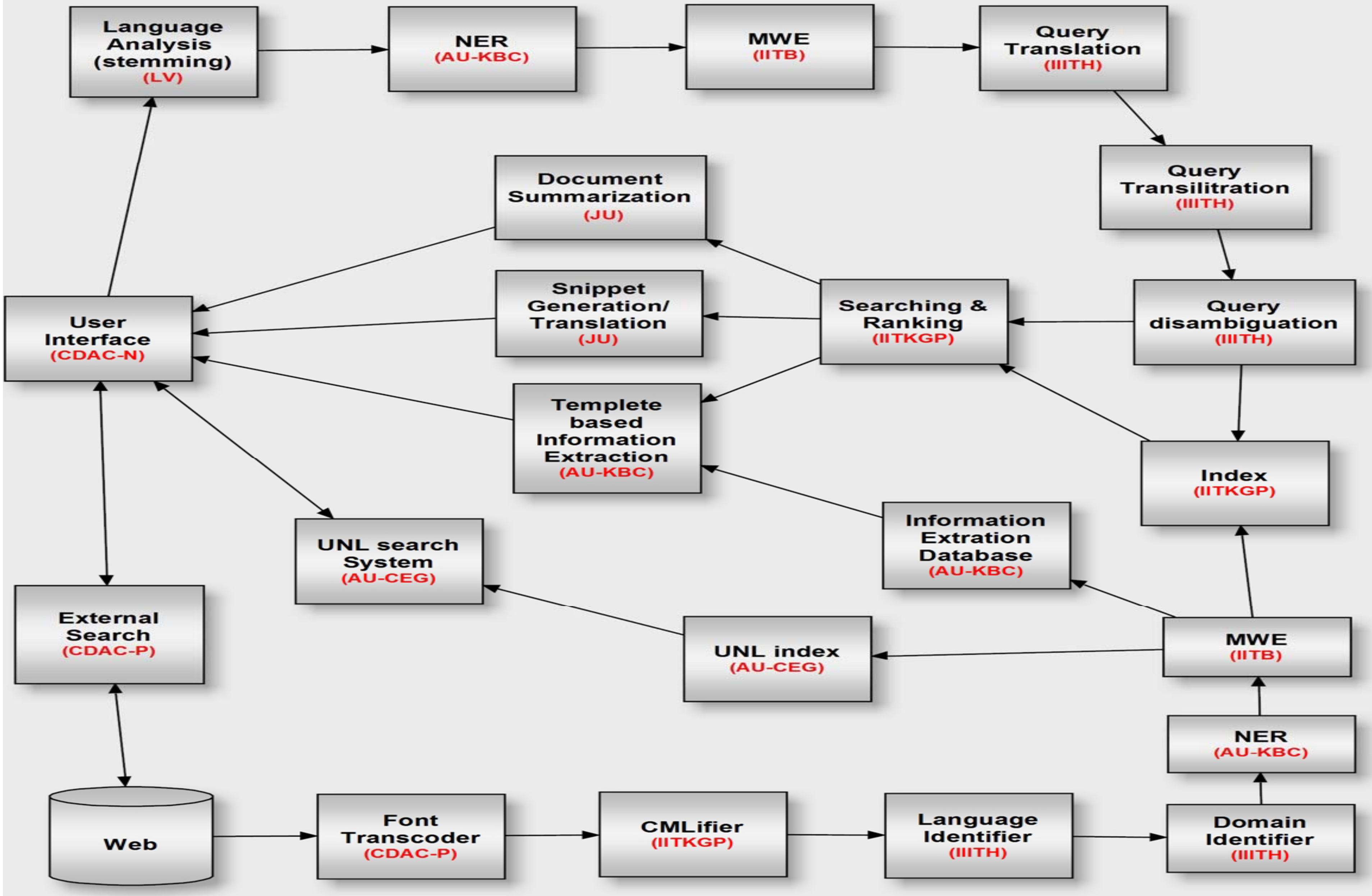
English Document

Snippet Generation and Translation

Cross Lingual Information Access

- Cross Lingual Information Access (CLIA)
  - A web portal supporting monolingual and cross lingual IR in 6 Indian languages and English
  - Domain : Tourism
  - It supports :
    - Summarization of web documents
    - Snippet translation into query language
    - Temple based information extraction
  - The CLIA system is publicly available at
    - http://www.clia.iitb.ac.in/clia-beta-ext

Cross Lingual Information Access



Language Analysis (stemming) (LV)

NER (AU-KBC)

MWE (IITB)

Query Translation (IIITH)

Query Transilitration (IIITH)

Document Summarization (JU)

Snippet Generation/ Translation (JU)

Searching & Ranking (IITKGP)

Query disambiguation (IIITH)

User Interface (CDAC-N)

Templete based Information Extraction (AU-KBC)

Index (IITKGP)

UNL search System (AU-CEG)

Information Extration Database (AU-KBC)

External Search (CDAC-P)

UNL index (AU-CEG)

MWE (IITB)

NER (AU-KBC)

Web

Font Transcoder (CDAC-P)

CMLifier (IITKGP)

Language Identifier (IIITH)

Domain Identifier (IIITH)

CLIA Demo

Various Ranking methods

- Vector Space Model
  - Lucene, Nutch , Lemur , etc
- Probabilistic Ranking Model
  - Classical spark John's ranking (Log ODD ratio)
  - Language Model
- Ranking using Machine Learning Algo
  - SVM, Learn to Rank, SVM-Map, etc
- Link analysis based Ranking
  - Page Rank, Hubs and Authorities, OPIC , etc

Nutch Ranking

- CLIA is built on top on Nutch – A open source web search engine.
- It is based on Vector space model

$$\text{[illegible]}(\text{[illegible]}, \text{[illegible]})$$

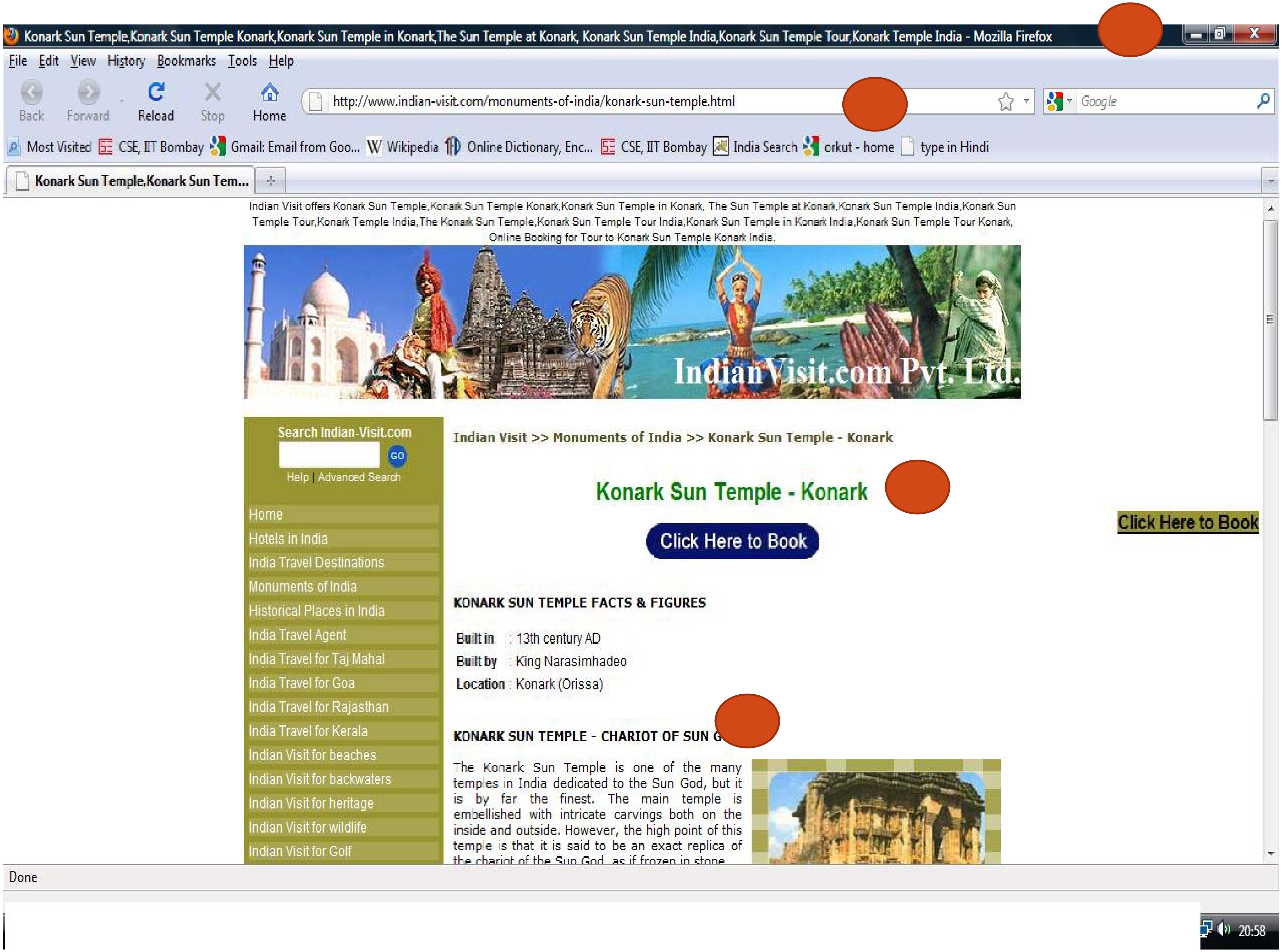
$$= \text{[illegible]}(\text{[illegible]}) * \text{[illegible]}(\text{[illegible]}, \text{[illegible]}) \sum_{\text{[illegible]}} \text{[illegible]}(\text{[illegible]}) * \text{[illegible]}(\text{[illegible]})^2$$

$$* \text{[illegible]}(\text{[illegible]})$$

$$\text{[illegible]}(\text{[illegible]}, \text{[illegible]}) = \frac{\overrightarrow{\text{[illegible]}} \; \overrightarrow{\text{[illegible]}}}{|\overrightarrow{\text{[illegible]}}||\overrightarrow{\text{[illegible]}}|}$$

Link analysis

- Calculates the importance of the pages using web graph
  - Node: pages
  - Edge: hyperlinks between pages
- Motivation: link analysis based score is hard to manipulate using spamming techniques
- Plays an important role in web IR scoring function
  - Page rank
  - Hub and Authority
  - Online Page Importance Computation (OPIC)
- Link analysis score is used along with the tf-idf based score
- We use OPIC score as a factor in CLIA.


Konark Sun Temple,Konark Sun Temple Konark,Konark Sun Temple in Konark,The Sun Temple at Konark, Konark Sun Temple India,Konark Sun Temple Tour,Konark Temple India - Mozilla Firefox
File Edit View History Bookmarks Tools Help
Back Forward Reload Stop Home
http://www.indian-visit.com/monuments-of-india/konark-sun-temple.html
Google
Most Visited CSE, IIT Bombay Gmail: Email from Goo... Wikipedia Online Dictionary, Enc... CSE, IIT Bombay India Search orkut - home type in Hindi
Konark Sun Temple,Konark Sun Tem...
IndianVisit.com Pvt. Ltd.
Search Indian-Visit.com
GO
Help | Advanced Search
Home
Hotels in India
India Travel Destinations
Monuments of India
Historical Places in India
India Travel Agent
India Travel for Taj Mahal
India Travel for Goa
India Travel for Rajasthan
India Travel for Kerala
Indian Visit for beaches
Indian Visit for backwaters
Indian Visit for heritage
Indian Visit for wildlife
Indian Visit for Golf
Click Here to Book
Done
20:58

Indian Visit offers Konark Sun Temple,Konark Sun Temple Konark,Konark Sun Temple in Konark, The Sun Temple at Konark,Konark Sun Temple India,Konark Sun Temple Tour,Konark Temple India,The Konark Sun Temple,Konark Sun Temple Tour India,Konark Sun Temple in Konark India,Konark Sun Temple Tour Konark, Online Booking for Tour to Konark Sun Temple Konark India.

Indian Visit >> Monuments of India >> Konark Sun Temple - Konark

Konark Sun Temple - Konark

Click Here to Book

KONARK SUN TEMPLE FACTS & FIGURES

Built in : 13th century AD
Built by : King Narasimhadeo
Location : Konark (Orissa)

KONARK SUN TEMPLE - CHARIOT OF SUN G

The Konark Sun Temple is one of the many temples in India dedicated to the Sun God, but it is by far the finest. The main temple is embellished with intricate carvings both on the inside and outside. However, the high point of this temple is that it is said to be an exact replica of the chariot of the Sun God, as if frozen in stone.

Learning a ranking function

- How much weight should be given to different part of the web documents while ranking the documents?
- A ranking function can be learned using following method
  - Machine learning algorithms: SVM, Max-entropy
  - Training
    - A set of query and its some relevant and non-relevant docs for each query
    - A set of features to capture the similarity of docs and query
    - In short, learn the optimal value of features
  - Ranking
    - Use a Trained model and generate score by combining different feature score for the documents set where query words appears
    - Sort the document by using score and display to user

Extended Features for Web IR

1. Content based features
   - Tf, IDF, length, co-ord, etc
2. Link analysis based features
   - OPIC score
   - Domains based OPIC score
3. Standard IR algorithm based features
   - BM25 score
   - Lucene score
   - LM based score
4. Language categories based features
   - Named Entity
   - Phrase based features

Content based Features

Feature	Formulation	Descriptions
C1	$\sum_{\square_{\square} \in \square \cap \square} \square(\square_{\square}, \square)$	Term frequency (tf)
C2	$\sum_{\square_{\square} \in \square \cap \square} \log(\square(\square_{\square}, \square) + 1)$	SIGIR feature
C3	$\sum_{\square_{\square} \in \square \cap \square} \frac{\square(\square_{\square}, \square)}{\lvert\square\rvert}$	Normalized tf
C4	$\sum_{\square_{\square} \in \square \cap \square} \log\left(1 + \frac{\square(\square_{\square}, \square)}{\lvert\square\rvert}\right)$	SIGIR feature
C5	$\sum_{\square_{\square} \in \square \cap \square} \log\left(\frac{\lvert\square\rvert}{\square\square(\square_{\square})}\right)$	Inverse doc frequency (IDF)
C6	$\sum_{\square_{\square} \in \square \cap \square} \log\left(\log\left(\frac{\lvert\square\rvert}{\square\square(\square_{\square})}\right)\right)$	SIGIR feature
C7	$\sum_{\square_{\square} \in \square \cap \square} \log\left(1 + \frac{\lvert\square\rvert}{\square(\square_{\square}, \square)}\right)$	SIGIR feature
C8	$\sum_{\square=1}^{\square} \log\left(1 + \frac{\square(\square_{\square}, \square)}{\lvert\square\rvert} \square\square\square(\square_{\square})\right)$	Tf*IDF
C9	$\sum_{\square_{\square} \in \square \cap \square} \log\left(1 + \frac{\square(\square_{\square}, \square)}{\lvert\square\rvert} \log \frac{\lvert\square\rvert}{\square\square(\square_{\square})}\right)$	SIGIR feature
C10	$\sum_{\square_{\square} \in \square \cap \square} \log\left(1 + \frac{\square(\square_{\square}, \square)}{\lvert\square\rvert} \frac{\lvert\square\rvert}{\square(\square_{\square}, \square)}\right)$	SIGIR feature

Details of features

Feature No	Descriptions
1	Length of body
2	length of title
3	length of URL
4	length of Anchor
5-14	C1-C10 for Title of the page
15-24	C1-C10 for Body of the page
25-34	C1-C10 for URL of the page
35-44	C1-C10 for Anchor of the page
45	OPIC score
46	Domain based classification score

Details of features(Cont)

Feature No	Descriptions
48	BM25 Score
49	Lucene score
50	Language Modeling score
51 -54	Named entity weight for title, body , anchor , url
55-58	Multi-word weight for title, body , anchor , url
59-62	Phrasal score for title, body , anchor , url
63-66	Co-ord factor for title, body , anchor , url
71	Co-ord factor for H1 tag of web document

Experiments and results

				MAP
Nutch Ranking	**0.2267**	**0.2267**	**0.2667**	**0.2137**
DIR with Title + content	0.6933	0.64	0.5911	0.3444
DIR with URL+ content	0.72	0.62	0.5333	0.3449
DIR with Title + URL + content	0.72	0.6533	0.56	0.36
DIR with Title+URL+content+anchor	0.73	0.66	0.58	0.3734
DIR with Title+URL+ content + anchor+ NE feature	**0.76**	**0.63**	**0.6**	**0.4**

Thanks