Insights On Information Retrieval System

Presented By:

Dhiraj Bashyal(MSIISE05)

Pitambar Khanal(MSIISE13)

Sarthak Pokharel(MSIISE18)

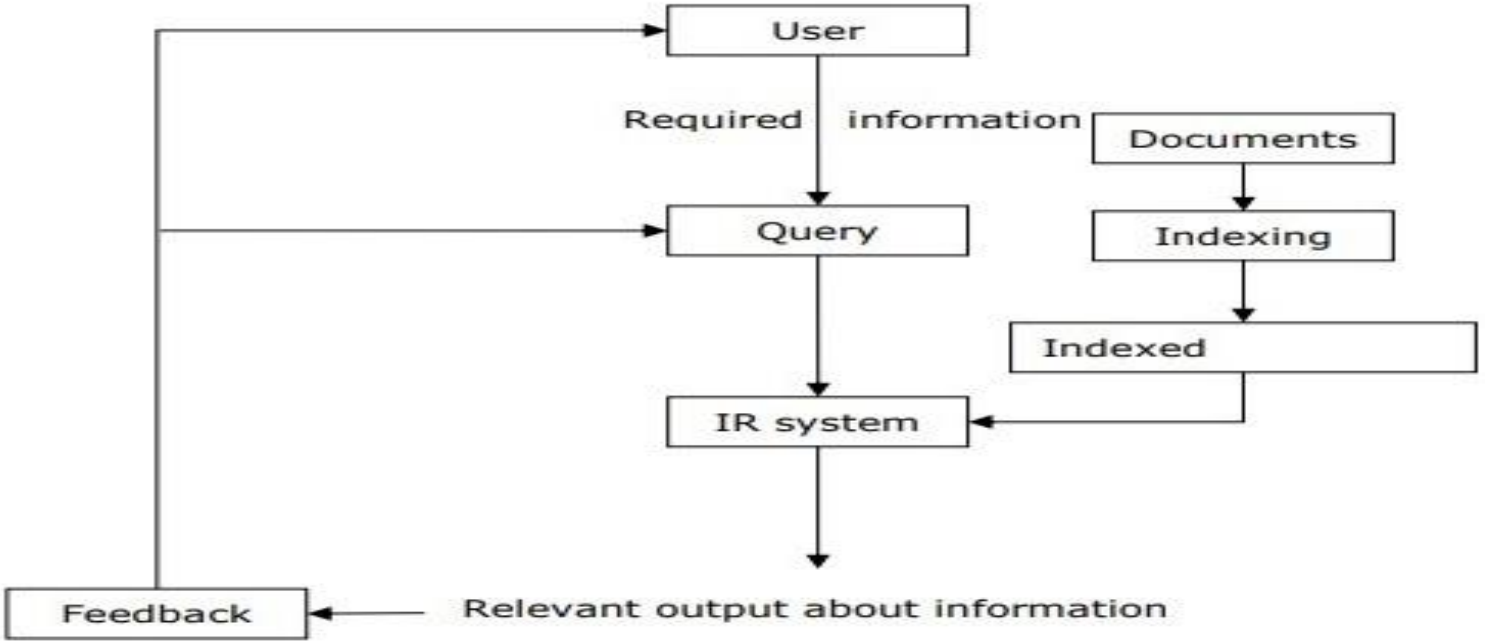
Outline

- Information Retrieval Basics
- Classical, Non Classical and Alternative Models of IR
- Relation Matching
- Conceptual Graphs in IR
- Cross lingual IR
- Evaluating IR Systems

Information Retrieval Basics[1]

- Software program that deals with the organization, storage, retrieval and evaluation of information:
  from document repositories particularly textual information
- Informs the existence and location of documents that might consist of the required information
- The documents that satisfy user’s requirement are called relevant documents.
- A perfect IR system will retrieve only relevant documents

Information Retrieval Basics[2]



Information Retrieval Basics[3]

Here are the prerequisites for an IR model:

1. An automated or manually-operated indexing system used to index and search techniques and procedures.
2. A collection of documents in any one of the following formats: text, image or multimedia.
3. A set of queries that serve as the input to a system, via a human or machine.
4. An evaluation metric to evaluate a system’s effectiveness (for instance, precision and recall)

Information Retrieval Basics[4]

The various components of an Information Retrieval Model include:

1.Acquisition

- Data is compiled by web crawlers and is sent to database storage systems.

2.Representation

- The free-text terms are indexed, and the vocabulary is sorted, both using automated or manual procedures
- For instance, a document abstract will contain a summary, meta description, bibliography, and details of the authors or co-authors

3.File Organization

- File organization is carried out in one of two methods, sequential or inverted
- Sequential file organization involves data contained in the document.
- The Inverted file comprises a list of records, in a term by term manner

Information Retrieval Basics[5]
Evaluation of Information Retrieval System[1]

- Evaluation of information retrieval system measure:
- Which of the two existing system perform better and try to assess how the level of performance of a given can be improved
- Effectiveness and Efficiency are two basic parameter for measuring the performance of system
- Effectiveness means measure of how far it can retrieve relevant information while withholding non-relevant information
- Efficiency means how economically the system is achieving its objectives

Information Retrieval Basics[6]

Evaluation of Information Retrieval System[2]

Lancaster state that evaluation of information retrieval system can be justified by the following three issues:

1. How well the system is satisfying its objectives
2. How efficiently it is satisfying its objectives
3. Whether the system justified its existence

Information Retrieval Basics[7]

Evaluation of Information Retrieval System[3]

To measure **information retrieval effectiveness** in the standard way, we need a test collection consisting of three things:

1. A document collection

2. A test suite of information needs, expressible as queries

3. A set of relevance judgments, standardly a binary assessment of either relevant or irrelevant for each query-document pair

Information Retrieval Basics[8]

Evaluation of Information Retrieval System[4]

EVALUATION CRITERIA

Evaluation of Information Retrieval is conduct into two different viewpoints

1. Managerial view:

- When evaluation is conducted from managerial point of view it is called managerial oriented evaluation

2. User view:

- When evaluation is conducted from the user point of view it is called user-oriented evaluation study

Information Retrieval Basics[9]

Cleverdon in 1966 identified **six criteria for the evaluation of an information retrieval system**

1. Recall-
   - Ability of the system to present all the relevant items

$$\text{Recall} = \frac{\textbf{Number of relevant item retrieved}}{\textbf{Total number of relevant items in the collection}} \times 100$$

2. Precision-
   - Ability of the system to present only those items that is relevant

$$\text{Precision} = \frac{\textbf{Number of relevant item retrieved}}{\textbf{Total number of items retrieved}} \times 100$$

Thus,recall related to the ability of the system to retrieve relevant documents and precision related to its ability not to retrieve non-relevant documents

Information Retrieval Basics[10]

3. Time lag-
   - Average interval between the time the search request is made and the time an answer is provided.

4. Effort
   - Intellectual as well as physical required from the user in obtaining answer to the search request

5. Form of presentation
   - Search output, which effects the user ability to make use of the relevant items

6. Coverage of the collection
   - The extent to which the system includes relevant matter

Types of Information Retrieval (IR) Model

An information retrieval model (IR) model can be classified into the following three models:

1. Classical IR Model
2. Non Classical IR Model
3. Alternative IR Model

Classical IR Model[1]

- Model based on mathematical knowledge that was easily recognized and understood as well
- Boolean, Vector and Probabilistic are the three classical IR models

Boolean Model

- Simple retrieval model based on set theory and Boolean algebra
- Boolean model provides a framework that is easy to grasp by a common user of an IR system
- Queries are specified as Boolean expressions which have precise semantics
- It only retrieves exact matches and gives the user, a sense of control over the system

Classical IR Model[2]

The Boolean model can be defined as −

- D − A set of words, i.e., the indexing terms present in a document. Here, each term is either present (1) or absent (0).
- Q − A Boolean expression, where terms are the index terms and operators are logical products − AND, logical sum − OR and logical difference − NOT
- F − Boolean algebra over sets of terms as well as over sets of documents If we talk about the relevance feedback, then in Boolean IR model the Relevance prediction can be defined as follows −
- R − A document is predicted as relevant to the query expression if and only if it satisfies the query expression as − $((text \vee information) \wedge rerie val \wedge \tilde{} theory)$

Classical IR Model[3]

Major drawbacks of Boolean model are:

- First, its retrieval strategy is based on a binary decision criterion
- Second,The query language is expressive, but it is complicated too

The Boolean Model: Example

Given the following three documents, Construct Term – document matrix

Boolean model for the query “gold silver truck”

• D1: “Shipment of gold damaged in a fire”
• D2: “Delivery of silver arrived in a silver truck”
• D3: “Shipment of gold arrived in a truck”

Classical IR Model[4]

Query” Gold Silver Truck”

	Arrive	Damage	Deliver	Fire	Gold	Silver	Ship	Truck
D1	0	1	0	1	1	0	1	0
D2	1	0	1	0	0	1	0	1
D3	1	0	0	0	1	0	1	1
Query	0	0	0	0	1	1	0	1

Classical IR Model[5]

Vector-Space Model[1]

- Recognizes that the use of binary weights is too limiting and proposes a framework in which partial matching is possible
- Accomplished by assigning non-binary weights to index terms in queries and in documents.
- These term weights are ultimately used to compute the degree of similarity between each document stored in the system and the user query.
- By sorting the retrieved documents in decreasing order, the vector model takes into consideration documents that match the query terms only partially

Classical IR Model[6]

Vector-Space Model[2]

To understand more about the vector Space model, you have to understand the following points:

1. In this model, the index representations (documents) and the queries are represented by vectors in a T dimensional Euclidean space.

2. T represents the number of distinct terms used in the documents

3. Each axis corresponds to one term.

4. Ranked list of documents ordered by similarity to the query

- Where the similarity between a query and a document is computed using a metric on the respective vector

5. The similarity measure of a document vector to a query vector is usually the cosine of the angle between them

Classical IR Model[6]

Vector-Space Model[3]

$$\text{similarity} = \cos(\theta) = \frac{\mathbf{A} \cdot \mathbf{B}}{\|\mathbf{A}\|\|\mathbf{B}\|} = \frac{\sum_{i=1}^{n} A_i B_i}{\sqrt{\sum_{i=1}^{n} A_i^2}\sqrt{\sum_{i=1}^{n} B_i^2}},$$

Classical IR Model[6]
Vector-Space Model[4]

Advantages of the vector model are:
(1) Its term-weighting scheme improves retrieval performance
(2) Its partial matching strategy allows retrieval of documents that approximate the query conditions
(3) Its cosine ranking formula sorts the documents according to their degree of similarity to the query

Non-Classical IR Model[1]

- It is completely opposite to classical IR model.
- Such kind of IR models are based on principles other than similarity, probability, Boolean operations
- Information logic model, situation theory model and interaction models are the examples of non-classical IR model

Non-Classical IR Model[1]

Situation Theory[1]

- Logical approach that is based on a theory of information, Situation Theory
- Provides a powerful arsenal of concepts, which is useful in modelling documents and queries for the purpose of IR
- Information can be stored and relayed in different forms including texts, images, audio and video called documents
- A document is about a query if the situation representing the document contains information about the situation representing the query

Alternative IR Model[1]

- Enhancement of classical IR model making use of some specific techniques from some other fields
- Cluster model, fuzzy model and latent semantic indexing (LSI) models are the example of alternative IR model.

Latent Semantic Indexing (LSI) Model[1]

- Latent Semantic Indexing is a technique that projects queries and documents into a space with “latent” semantic dimensions
- A better approach that allow users to retrieve information on the basis of a conceptual topic or meaning of a document
- Tries to overcome the problems of lexical matching by using statistically derived conceptual indices

Alternative IR Model[2]

Latent Semantic Indexing (LSI) Model[2]

- LSI model does’t use individual words for retrieval
- LSI assumes that there is some underlying or latent structure in word usage that is partially obscured by variability in word choice
- A truncated SVD is used to estimate the structure in word usage across documents
- Retrieval is then performed using the database of singular values and vectors obtained from the truncated SVD
- *The latent semantic space that we project into has fewer dimensions than the original space*

Latent Semantic Indexing (LSI) Model[3]

- A dimensionality reduction technique takes a set of objects that exist in a high-dimensional space
- Represents them in a low dimensional space, often in a two-dimensional or three-dimensional space for the purpose of visualization
- SVD finds the optimal projection to a low dimensional space is the key property for exploiting word co-occurrence patterns
- For purposes of information retrieval, a user's query must be represented as a vector in k-dimensional space and compared to documents

Latent Semantic Indexing (LSI) Model[4]

- This method works by identifying the hidden contextual relationships between words:
  a. Latent → Hidden
  b. Semantic → Relationships Between Words
  c. Indexing → Information Retrieval
- When analysing a string of words, LSI removes conjunctions, pronouns, and common verbs, also known as stop words
- Isolates the words which comprise the main ‘content’ of a phrase

~~The~~ quick brown fox jumps ~~over the~~ lazy dog

Latent Semantic Indexing (LSI) Model[5]

- These words are then placed in a **Term Document Matrix** (TDM)
- TDM is a 2D grid that lists the frequency that each specific word occurs in the documents within a data set
- When words occur with the same general frequency in any documents, it is called **co-occurrence**

Latent Semantic Indexing (LSI) Model[6]

	Quick	Brown	Fox	Jumps	Over	Lazy	Dog
The quick brown fox jumps over the lazy dog	1	1	1	1	1	1	1
If the fox is quick he can jump over the dog.	1	0	1	0	1	0	1
Foxes are quick. Dogs are lazy.	0	1	1	0	0	1	1
Can a fox jump over a dog?	0	0	1	1	1	0	1

Latent Semantic Indexing (LSI) Model[7]

Advantages

1. True (latent) dimensions
   - LSI analysis recovers the original semantic structure of the space and its original dimension
2. Synonymy
   - Synonymy refers to the fact that the same underlying concept can be described using different terms
   - In LSI, the concept in question as well as all documents that are related to it are all likely to be represented by a similar weighted combination of indexing variables.

Latent Semantic Indexing (LSI) Model[8]

3. Polysemy

- Polysemy describes words that have more than one meaning, which is common property of language
- Large numbers of polysemous words in the query can reduce the precision of a search significantly

Disadvantages

1. Storage
2. Efficiency
3) Toward a theoretical foundation

Relation Matching[1]

- IR researchers in the late 1980s and early 1990s that the maximum attainable retrieval performance using keyword matching methods
- two ways in which relations can improve information retrieval effectiveness – through **relation matching** and through **query expansion**

Relation Matching[2]

- Relation matching
  - Both the **concepts** and the **relations** between the concepts as expressed in the **user's query are matched** with concepts and relations in documents
  - system first performs **concept or word matching**
- Query expansion
  - query concepts that are found in the document, the system further checks whether **the relations expressed between the query concepts match the relations expressed between the document concepts.**

Relation Matching[3] - Example

Consider the sentence Harry loves Sally.

If we use the sentence as a query in a keyword matching system, the system would look for documents containing the terms **Harry*, Sally* and love*** (where "*" is the truncation sign),

(1) Harry loves Sally.

(2) Sally loves Harry, but Harry hates Sally.

(3) Harry’s best friend loves Sally’s best friend.

(4) Harry and Sally loves pizza.

(5) Harry’s love for Sally is beyond doubt.

Relation Matching[4] - Example

- For example sentences from previous slide, sentences (1) and (5) should be ranked higher than the other sentences.
  - With relation matching, a document that not only has the keywords Harry, Sally and love but also expresses **the correct relation between the concepts** would be given a higher ranking in the retrieval results
- Relation matching improves **retrieval precision by reducing the number of non-relevant documents** retrieved
- System uses the additional criteria of relation matches to eliminate some non-relevant documents that would otherwise be retrieved by keyword matching

Relation Matching[5] - Example

Consider the sentence Harry loves Sally.

If we use the sentence as a query in a keyword matching system, the system would look for documents containing the terms **Harry*, Sally* and love*** (where "*" is the truncation sign),

Main issues and research questions relating to relation matching[1]

- Comparison with keyword matching
  - To what extent does the use of relation matching improve retrieval effectiveness compared with keyword matching alone
- Comparison with word proximity matching
  - To what extent does the use of relation matching improve retrieval results compared with using word proximity information?
- The difficulty of identifying relations automatically, especially when the database is not limited to a narrow subject area
  - How can the automatic identification of relations in a heterogeneous textual database be improved?
  - Are simple methods of identifying relations (and the relatively low accuracy) good enough to yield a material improvement in retrieval effectiveness?
  - Will more accurate identification of relations yield better retrieval results than simple methods?

Main issues and research questions relating to relation matching[2]

- The relation matching method
  - There are several ways of identifying relations automatically and several ways of performing relation matching.
  - There are also different types of relations and different sets of relations used by different researchers
- The method of combining relation with keyword matching
  - What is the relative importance of keyword and relation matches in information retrieval?
  - How can relation matching be combined with keyword matching to estimate the likelihood that a document is relevant to the user?
  - Should different types of relations be weighted differently?

Main issues and research questions relating to relation matching[3]

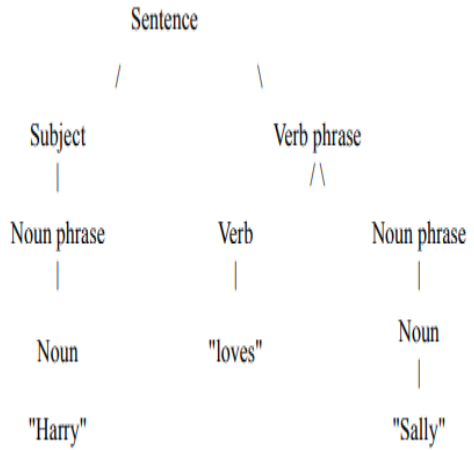
- The circumstances in which relation matching is important. Relation matching may not be helpful in all situations.
  - For what types of queries, documents, subject areas and applications is relation matching helpful in improving retrieval results?
  - How can a retrieval system be designed to identify the instances when relation matching is likely to improve the retrieval results?

Matching With Manually Identified Relations

- Some **human indexer** has indicated that there is a relationship between two or more concepts in the content of the document
- Use of relations in manual indexing has not caught on probably because its effectiveness has not been clearly demonstrated
- **The effective use of relations requires training both for the indexer and the user**

Matching With Automatically Identified Relations

- Syntactic relations refer to relations that are derived from the syntactic structure of the sentence.
- Determining the syntactic structure of a sentence is one of the processing steps needed for determining the semantic relations and the meaning of the sentence



Factors that affect the usefulness of relation matching

- the accuracy of the automatic identification of relations
- the method used for calculating the retrieval scores (e.g. a tree matching method or term matching method) .
- the type of documents and the type of queries.
- the type of relations used and the set of relations used (i.e. syntactic or semantic, and which particular relations?)
- the degree of relational ambiguity between the concepts linked by a relation

Information Retrieval with Conceptual Graph Matching[1]

- This representation incorporates the information about both the concepts mentioned in the text and their relationships

  e.g., [binary] ->(attr) -> [search]

  consider relations from a few basic types, such as attribute, subject, object, etc

- **Text mining**, such representations are used because they are easily extracted from texts and easily analyzed

Information Retrieval with Conceptual Graph Matching[2]

- a phrase **John loves Mary** is represented with a graph like

  [John] ← (subj) ← [love] → (obj) → [Mary], and not like [John] ← (love) → [Mary].

- Algebraic formulation of flow diagrams,

  The semantic analyzer generates one or more conceptual graphs out of such syntactic structure:[3]

  [algebraic] ← (attr) ← [formulation] → (of) → [flow-diagram:*]

Information Retrieval with Conceptual Graph Matching[3] Comparison of conceptual graphs

- In general terms, our algorithm for the comparison of two conceptual graph representations of two texts consists of two main parts:
  - Find the intersection of the two (set of) graphs,
  - Measure the similarity between the two (set of) graphs as the relative size of each one of their intersection graphs

Information Retrieval with Conceptual Graph Matching[4]
Comparison of conceptual graphs

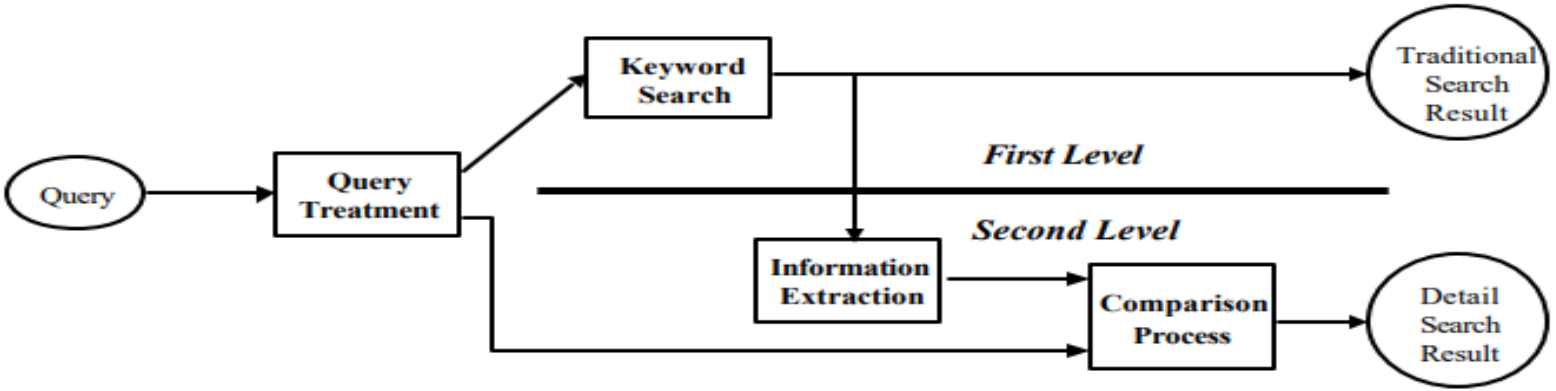
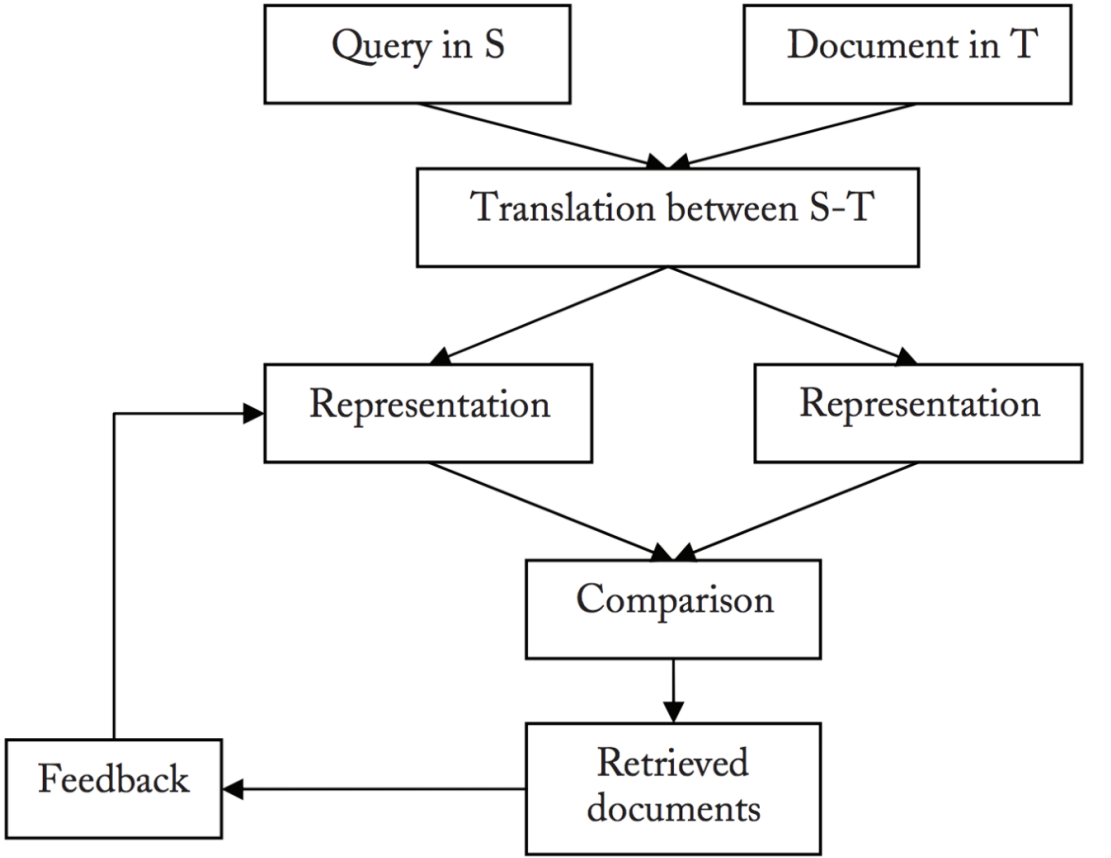


Fig. 3. Calculation of relational similarity.

Cross-lingual IR

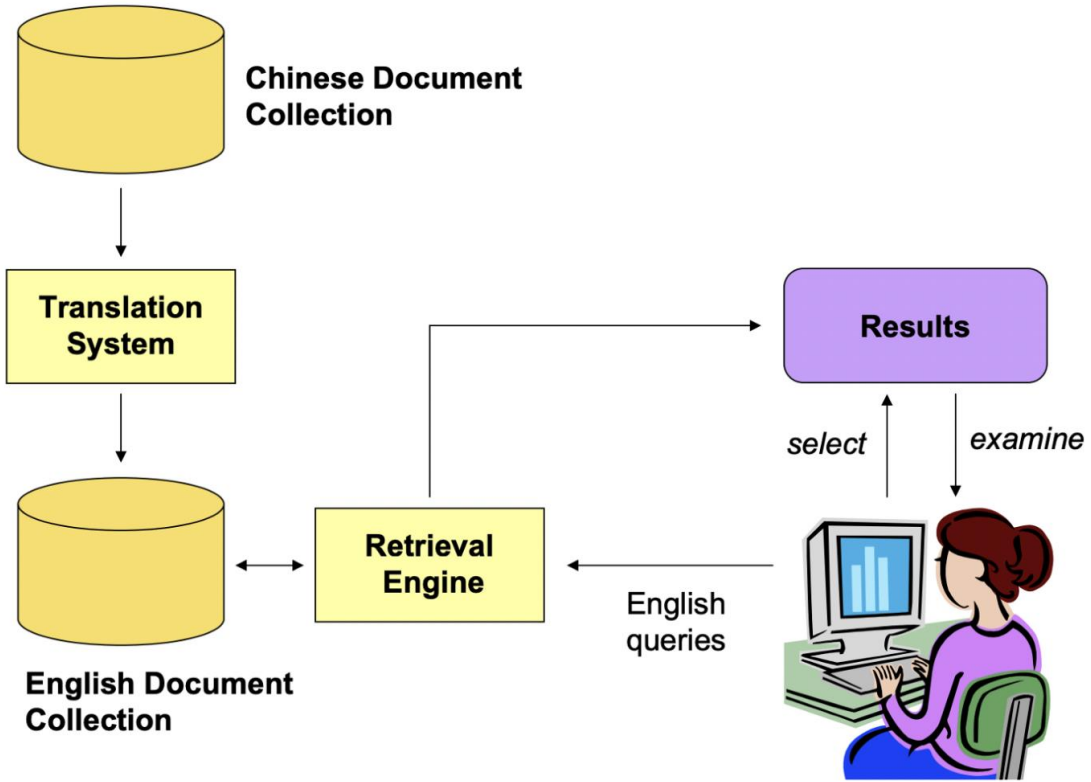
- Task of retrieving relevant information when the document is written in different language
- User search queries are not always answered in their native search languages
- A lot of information may not be present in the queried language.


Query in S
Document in T
Translation between S-T
Representation
Representation
Comparison
Retrieved documents
Feedback

Approaches to CLIR

1. **Query translation approach**
- The search query is translated into the language of document
- The user should be able to read/interpret the target language
- Translation can be done in word or phrase level
1. **Document translation approach**
- Translate the document to the queried language
- Lots of document to be translated
- Practically unscalable
1. **Interlingua based approach**
- Both documents and query are translated to common interlingua(like UNL)
- Requires huge resource as the translation is to be done online.

(Need to write answer in animation - Nishan, helppp !!!!!)



Challenges in CLIR

1. Translation ambiguity
2. Phrase identification and translation
3. Translate/Transliterate a term
4. Transliteration errors
5. Dictionary coverage
6. Font
7. Morphological analysis
8. Out of vocabulary (OOV) problems

Factors affecting performance of CLIR systems

1. **Limited size of dictionary**

- Leads to translational errors
- New words gets added frequently, difficult to timely update the dictionary
- Compounds and phrases cannot be contained completely in a dictionary

1. **Query translation / transliteration performance**

- Lexical ambiguity in both queried and translated languages.
- Increases irrelevant search key senses
- Extraneous senses can be added in the query language, translation alternatives can also have many senses.

Related works in CLIR

1. **Incremental crawling**

- Periodic crawling crawls a certain amount of page till certain depth and stops crawling. Recrawls same set of page after certain amount of time and replaces old crawls
- Incremental crawling keeps on crawling page refreshing the existing crawl and refreshing old pages with new ones
- Periodic crawl can index a new page only when the next crawling cycle begins while incremental crawling updates pages as soon as it is found
- Study of changes in web pages is important for incremental crawling.

Related works in CLIR - 2

2. Resource constraint crawling

- Periodic and incremental crawling requires lots of resources, not always available in small organizations.
- Throughput in crawling may be seen, crawling important url first might be beneficial.
- Classifying urls based on page contents.

3. Query translation and transliteration

- Translating the entire document into query language is less desirable due to enormous resource requirements
- Mostly query translation is used.
- Used in researches in different Hindi languages, for eg: use of SMT system, script representation which is common in all Hindi languages.

Evaluation of IR systems

- Traditionally, most important form of evaluation was to measure system effectiveness
- Currently, other user related metrics are introduced such as the user and the user environment
- Most effective form of evaluation will be the combination of the system effectiveness(how well the query is searched) and also the user environment
- Efficiency and Effectiveness
- Precision and recall, most common form of IR system evaluation

Evaluation using test collection - Cranfield approach

- Originated from experiments conducted in Cranfield library from 1985-1966
- Uses test collections : reusable and standardised resources that can be used to evaluate IR systems w.r.t system
- Focuses on comparative evaluation - compare various retrieval strategies or systems.
- Absolute score of system effectiveness is not preferred.

Cranfield approach (2)

1. Select different retrieval strategies to compare.
2. Use these to produce ranked list of documents(often called runs) for each query(often called topics)
3. Compute the effectiveness of each strategy for every query in the test collection as a function of relevant documents retrieved
4. Average the scores over all queries to compute overall effectiveness of the strategy or system
5. Use the scores to rank these strategies/systems relative to one another

Test collection

- Consists of 3 things:

a. document collection,
b. a set of topics that describes a user’s information need and
c. a set of relevance judgements indicating which documents in the collection are relevant to each topic

Document collection

- How many items should be gathered?
- What items should be sampled to create the document collection?
- What about copyright constraints?

Topics

- How should a suitable sets of topics be generated?
- How many topics are required for obtaining reliable evaluation results?
- Do the topic represent a diverse enough set of information needs?
- How should the topics be addressed?

Relevance assessments

- Who should do the assessments?
- How many assessments should be made?
- What are the assessors expected to do?
- What about finding missing relevant documents?

References

[1] G. Salton. The SMART Retrieval System. Englewood Cliffs, N.J., 1971. Prentice Hall, Inc.

[2] C.J. van Rijsbergen. Information Retrieval. Butterworths, second edition, 1979.

[3] Chris H. Q. Ding, A Similarity-Based Probability Model for Latent Semantic Indexing, Proc. of SIGIR'99, Berkeley, August 1999

[4] Levow, G.-A., Oard, D. W., and Resnik, P., Dictionary-based techniques for cross-language information retrieval, Information Processing and Management, 41, 523–547, 2005.

[5] Khoo, Christopher Soo-Guan (1997). "The Use of Relation Matching in Information Retrieval". Electronic Journal ISSN 1058-6768, September 1997

Thank you !!!