Inference Networks for Information Retrieval

Sumit Bhatia

Outline

- Introduction
- Inference Network based Retrieval Model
- Experimental Results
- Conclusions

Introduction

What are the important characteristics of a retrieval model?

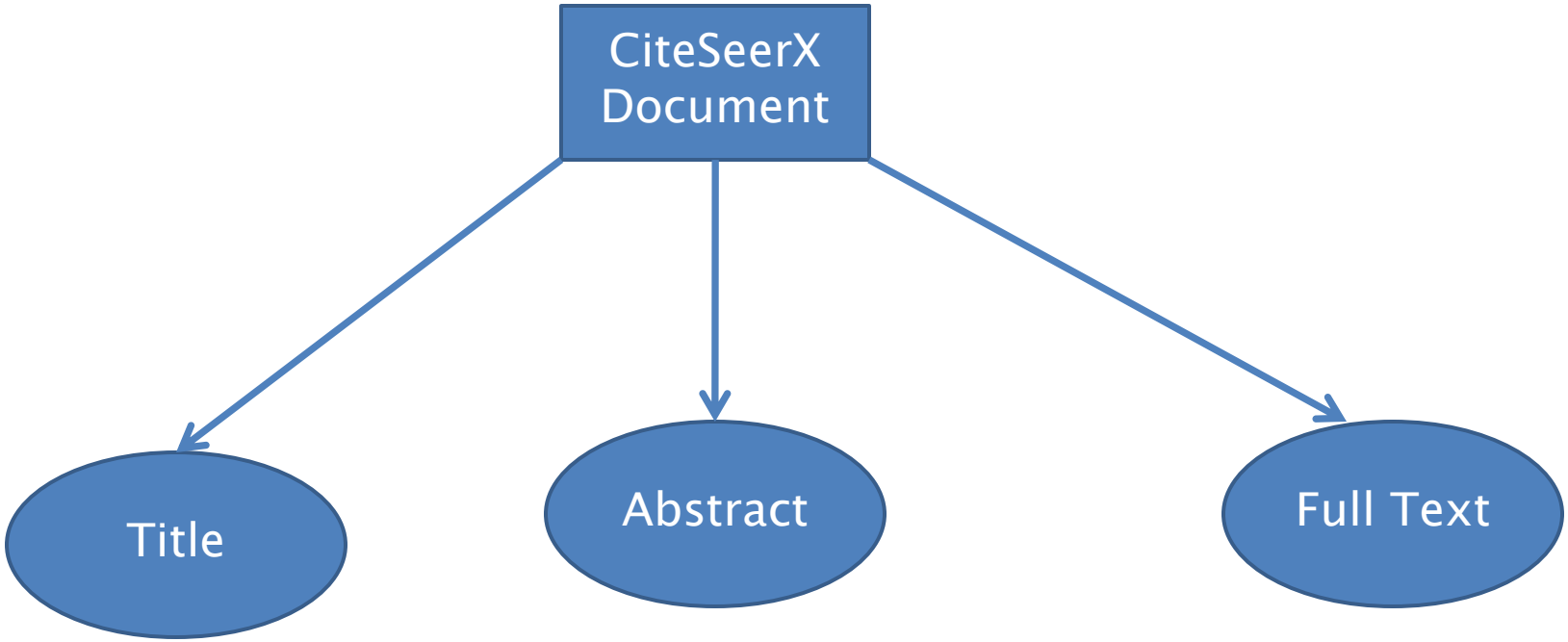
- Document Representation
- Query Representation
- Matching Function

Introduction

- Probabilistic Ranking Principle
  - P(R=1|d,q)
  - 1-0 Loss model, BM25 Model
- Language Models
  - P(d|q)
  - Query Likelihood
  - Relevance Based LM
  - Translation Models
  - Query document divergence

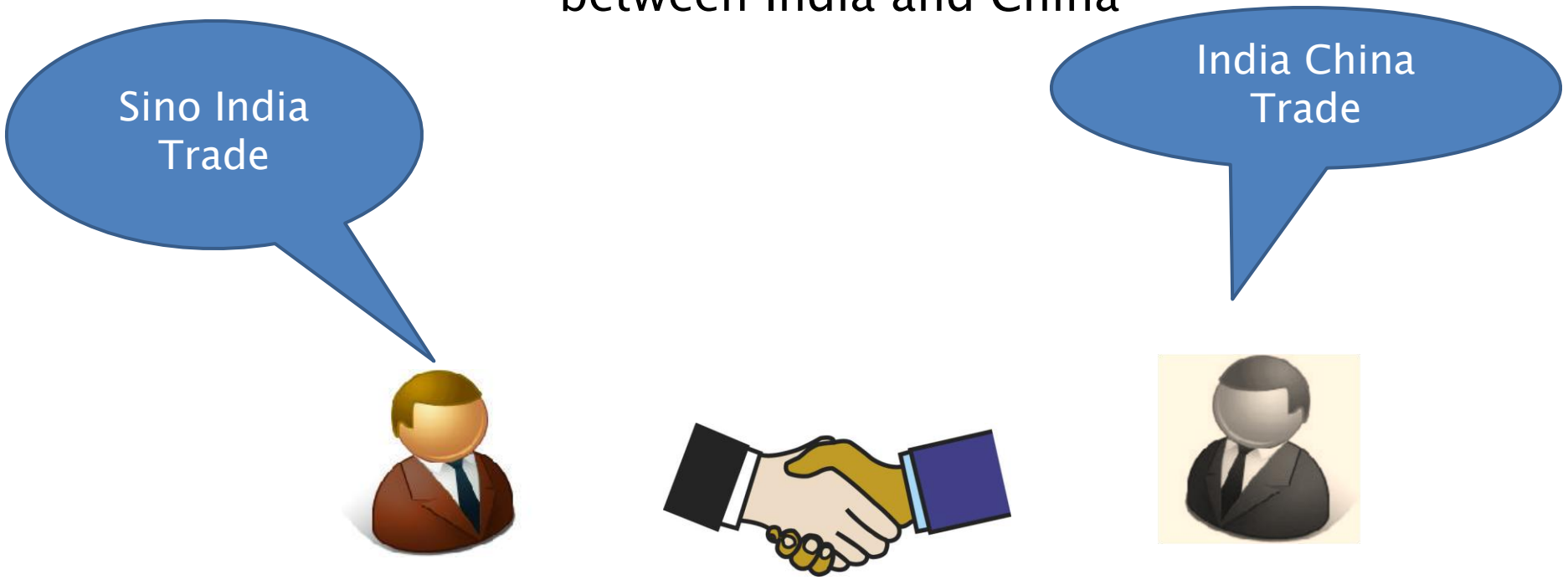
Motivation

- Utilize multiple document representation schemes



Motivation

Information Need : To find about trade relationships between India and China



How to combine these results???

Motivation

Nittany Lions vs Buckeyes
Or
Penn State vs Ohio State

- Different vocabularies of document author and query author
- Mapping between related concepts

Inference Network for IR

- Turtle and Croft (1991) introduced a retrieval model based on Inference Networks

- Described IR as an inference process
  - Users has information need I
  - Expressed through a query Q
  - A document is “observed”
  - Estimate the probability that user’s information need is satisfied by the document
  - Different query/document representations provide the evidence

Inference Network for IR

Inference Network (Inet)

- Directed Acyclic Graph

- Each node represents an **event** with a continuous /discrete outcome
  - In this talk, we assume binary outputs

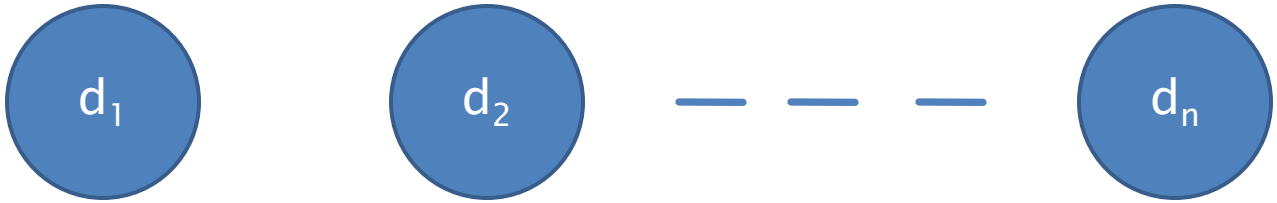
- Root node event occurs with a **prior** probability

- Non root nodes store a table of conditional probabilities aka **Link Matrix** describing its outcome given parent node

Inference Network for IR

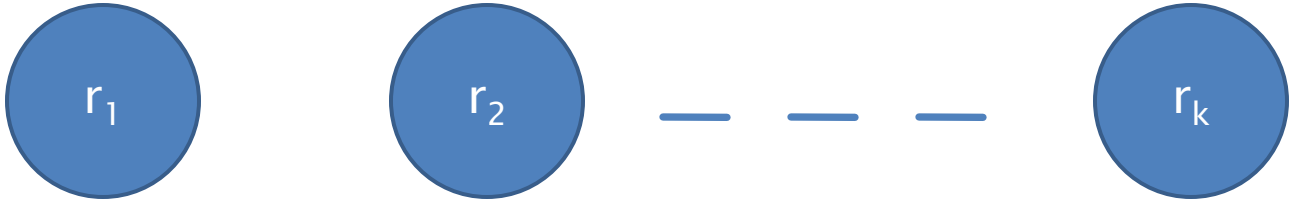
- 4 types of nodes:
  - Document Nodes
  - Representation Nodes
  - Query Representation Nodes
  - Information Need Node

Document Nodes



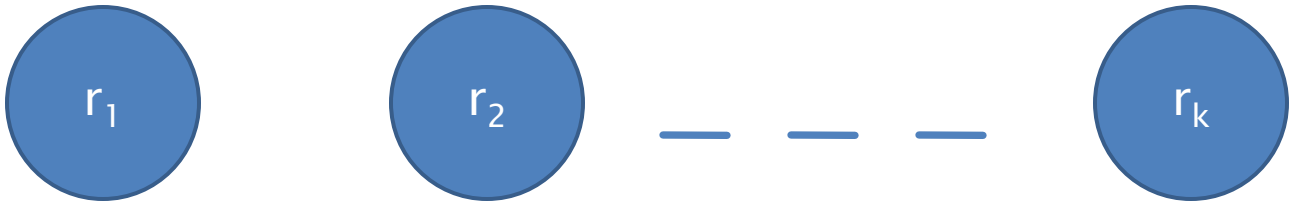
- One node for each document in the corpus
- Corresponds to the event that document d_i is observed
- Only one document observed at a point

Representation Nodes



- A set of representation nodes
- Representation = an indexable feature of document, eg. terms, phrases etc.
- Each document node acts as a parent for representation nodes

Representation Nodes



Since only one document is observed at a time, we wish to estimate the following for representation nodes:

$$P(r_k = true | d_j = true, d_{i \neq j} = false)$$

Also known as belief, bel(r_k)

Query Nodes

- Combine beliefs about representation nodes and other query nodes
- Enable structured query operators
- Need to compute Link Matrices

Query Nodes

- Suppose query node *q* has *n* parent nodes with corresponding beliefs p_1, p_2, …, p_n
- There are 2^n possible parent states and 2 possible outcomes of *q*
- Represented by a 2 x 2^n Link Matrix

$$L_{and} = \begin{pmatrix} 1 & 1 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}$$

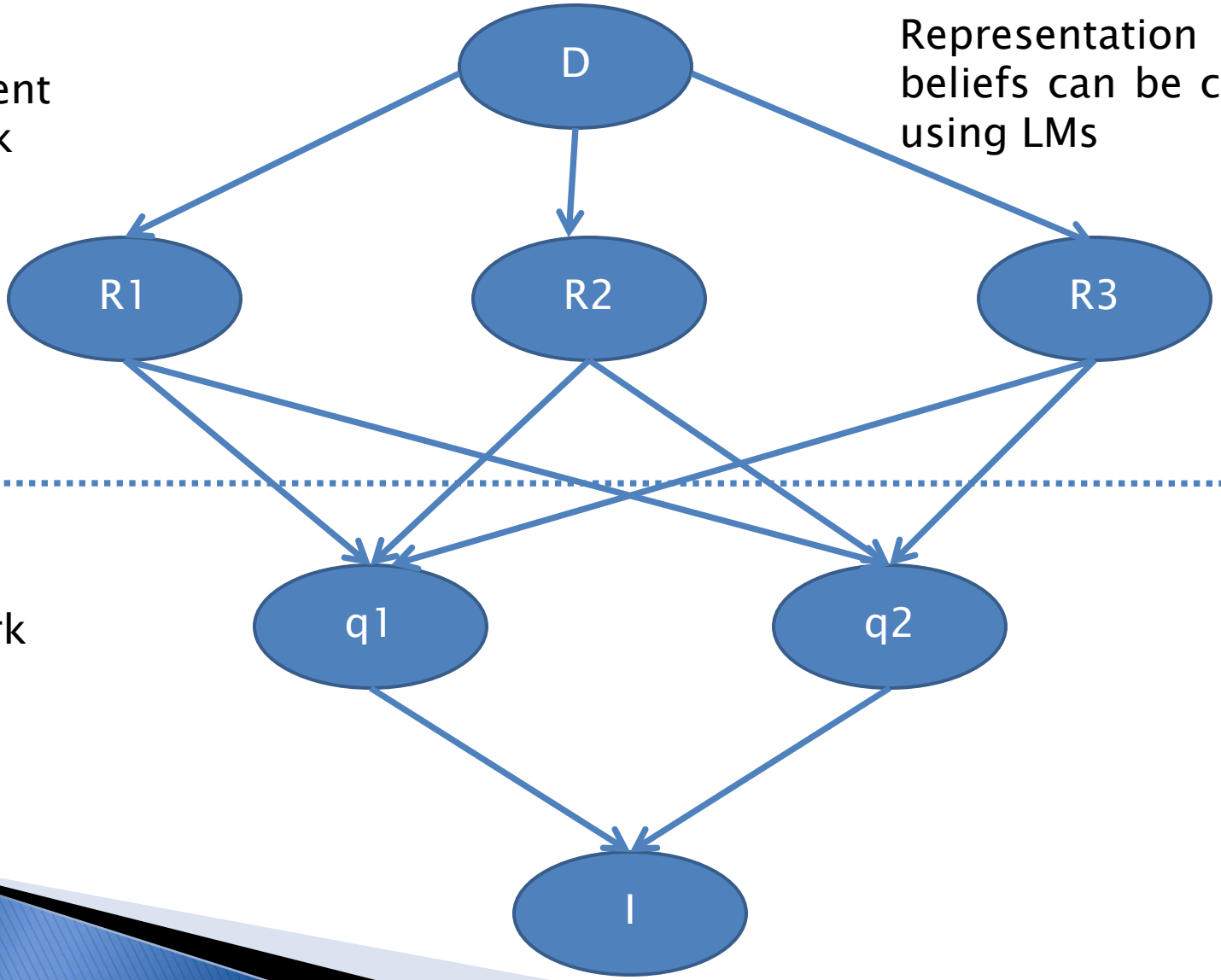
Information Need Node

- Represents the event that the information need of user is met
- Combines evidence from all other query nodes into a single belief
- Documents are ranked by this belief score

Putting All Pieces Together

Document Network

Representation Node beliefs can be computed using LMs



Query Network

Inets – Performance (Metzler and Croft, 2004)

- Trec 4,6,7, and 8 datasets
- For Trec 4, structured queries were available
- For Trec 6,7 and 8 structured queries were generated from title and description fields
- Example structured query for free text query “capital punishment deterrent crime”

```
#WAND (1.0 #WAND ( 1.0 capital 1.0 punishment
                   1.0 deterrent 1.0 crime
                   2.0 #UW20(capital punishment deterrent)
                   1.0 #PHRASE(capital punishment)
       1.0 #PASSAGE200 ( 1.0 capital 1.0 punishment
                         1.0 deterrent 1.0 crime
                         1.0 #PHRASE(capital punishment) ) )
```

Inets – Performance (Metzler and Croft, 2004)

	QL		InQuery		StructLM	
	Auto	Manual	Auto	Manual	Auto	Manual
Rel	6086	6086	6086	6086	6086	6086
Rret	3190	3371	3306	3679	3355	3737
0.0	0.6761	0.7166	0.7188	0.7896	0.6888	0.7893
0.1	0.4796	0.5082	0.4944	0.5601	0.4983	0.5479
0.2	0.3801	0.4156	0.3942	0.4581	0.3926	0.4486
0.3	0.3220	0.3465	0.3310	0.3883	0.3300	0.3997
0.4	0.2672	0.2960	0.2844	0.3317	0.2769	0.3329
0.5	0.2087	0.2315	0.2241	0.2552	0.2268	0.2631
0.6	0.1546	0.1708	0.1622	0.1849	0.1713	0.2079
0.7	0.0903	0.1033	0.0975	0.1236	0.1331	0.1520
0.8	0.0480	0.0567	0.0544	0.0727	0.0763	0.0894
0.9	0.0056	0.0175	0.0194	0.0300	0.0246	0.0422
1.0	0.0021	0.0038	0.0016	0.0014	0.0048	0.0024
Avg	0.2179	0.2397	0.2312	0.2688	0.2376	0.2779

Indri – Inets in Action

- Part of Lemur Toolkit
- Based on Inference Networks and LMs
- Powerful query language to support structured queries and evidence combination
- Different document types – TREC style, XML, HTML, .doc
- XML retrieval, sentence retrieval, passage retrieval, extent retrieval
- Scalable

Conclusion

- Information Retrieval can be considered as an inference process
- Inference Networks provide a theoretically sound framework to infer the probability that a given document satisfies user's information needs
- Language Models provide a means to estimate the beliefs/probabilities

References

1. D. Metzler and W.B. Croft. **Combining the language model and inference network approaches to retrieval.** Info. Proc. and Mgt. 40(5):735-750, 2004.
2. Lemur Project Website: http://www.lemurproject.org
3. Trevor Strohman, Donald Metzler, Howard Turtle and W. Bruce Croft. Indri: A language-model based search engine for complex queries. Umass Technical Report
4. H. Turtle and W. B. Croft, Evaluation of an inference network based retrieval model. Trans. Inf. Syst., 9(3):187-22, 1991.

Questions???