

The INQUERY Retrieval System

James Callan, Bruce Croft
and Stephen Harding
University of Massachusetts

Presented by Alex Ten,
CS 541, November 9.

System's goals

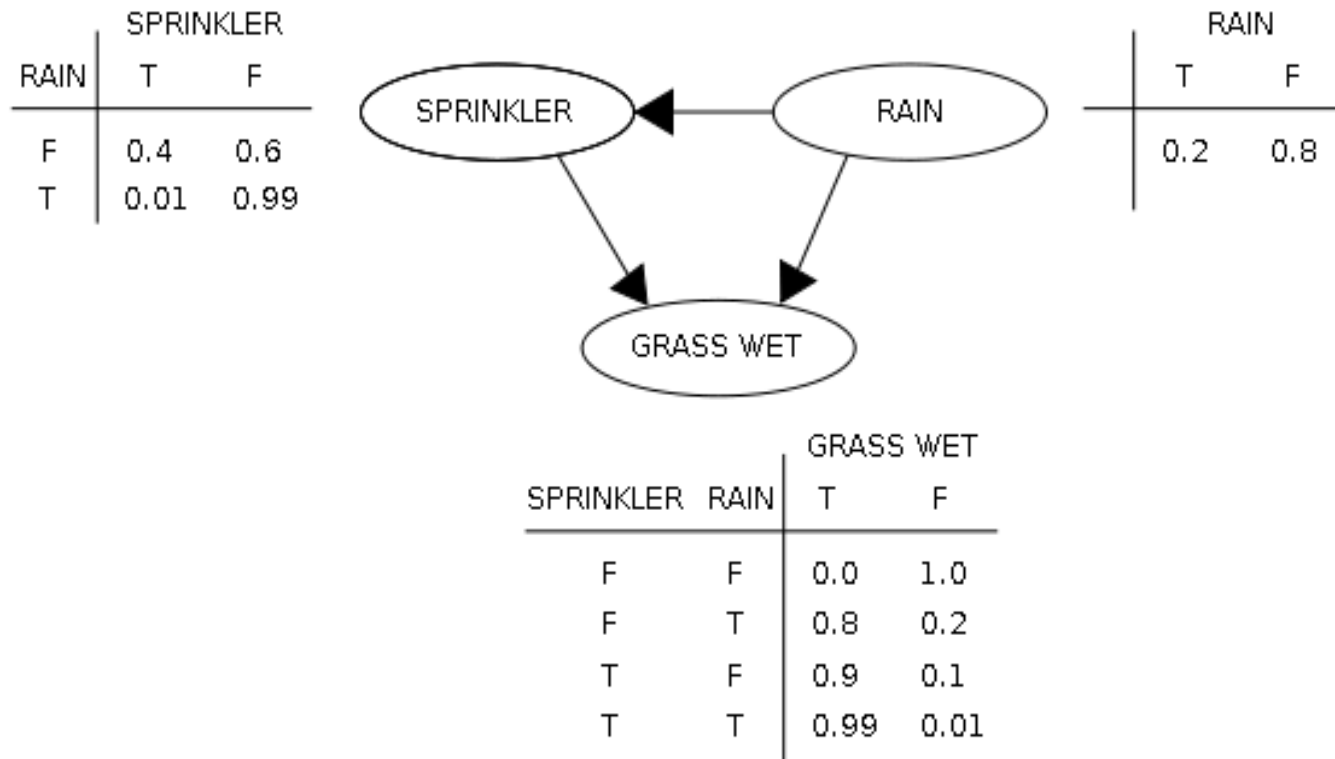
- Efficiency
- Complex queries
- Flexibility

Theory

The INQUERY system is based on a form of probabilistic retrieval model called the Bayesian inference network.

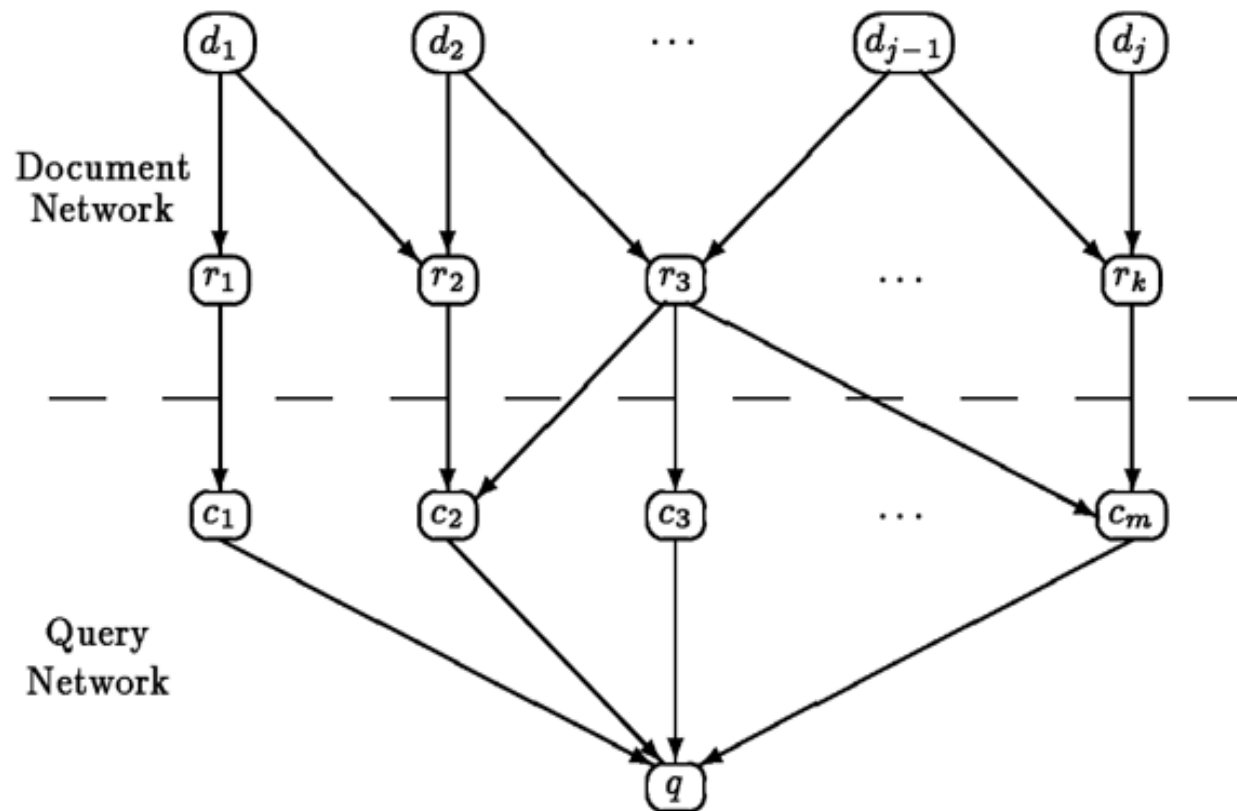
A Bayesian inference network is a directed acyclic graph in which nodes represent propositional variables and arcs represent dependencies.

A simple Bayes network

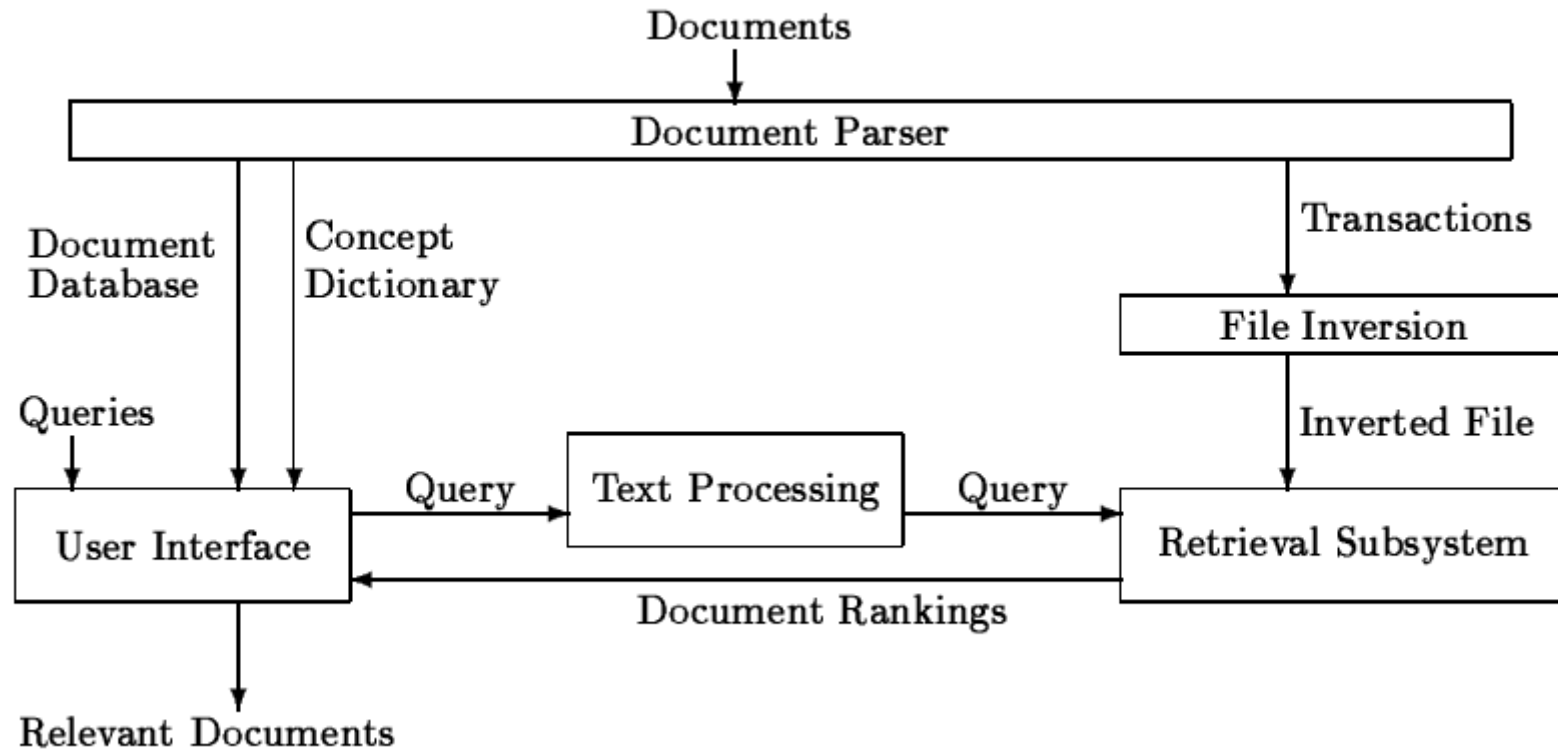


This picture is property of AnAj.

A simple Bayes network used in INQUERY



INQUERY Architecture



Document parser

- Lexical analyzer
- Syntactic analyzer
- Concept recognizer
- Transaction generator

Example

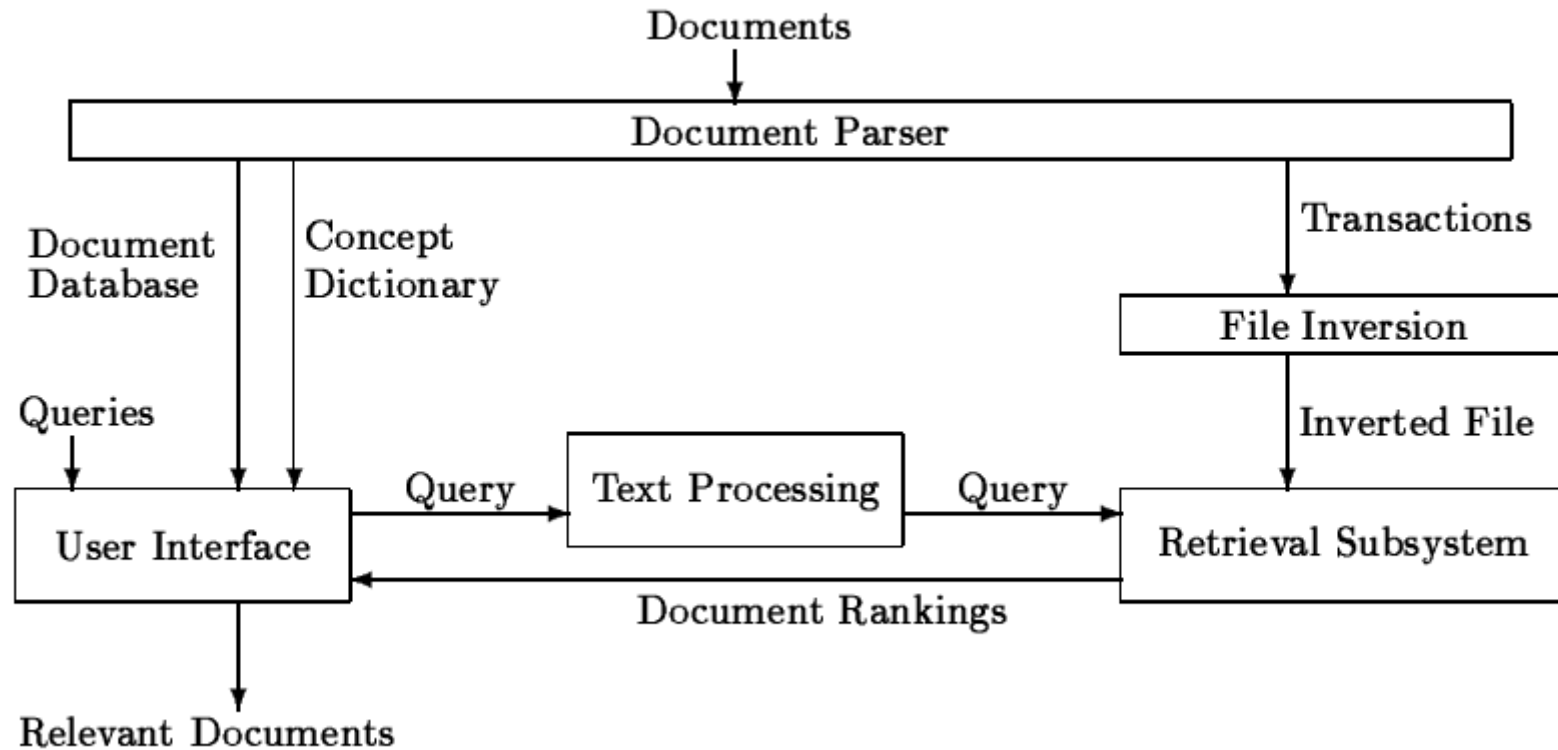
The person name recognizer uses a strategy similar to the company name recognizer, except that it looks for occupation and honorific titles.

Underlined words are now tokens.

Document parser

- Lexical analyzer
- Syntactic analyzer
- Concept recognizer
- Transaction generator

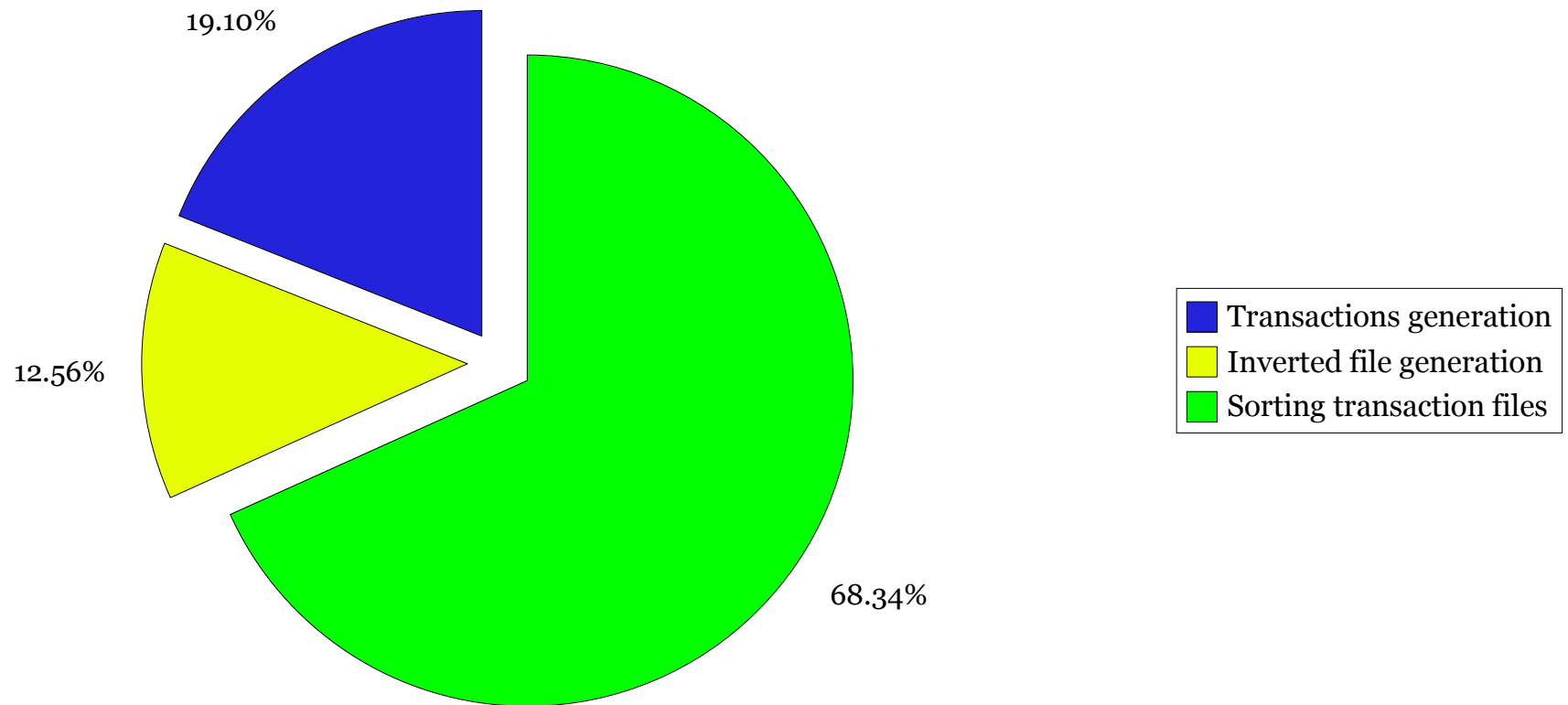
INQUERY Architecture



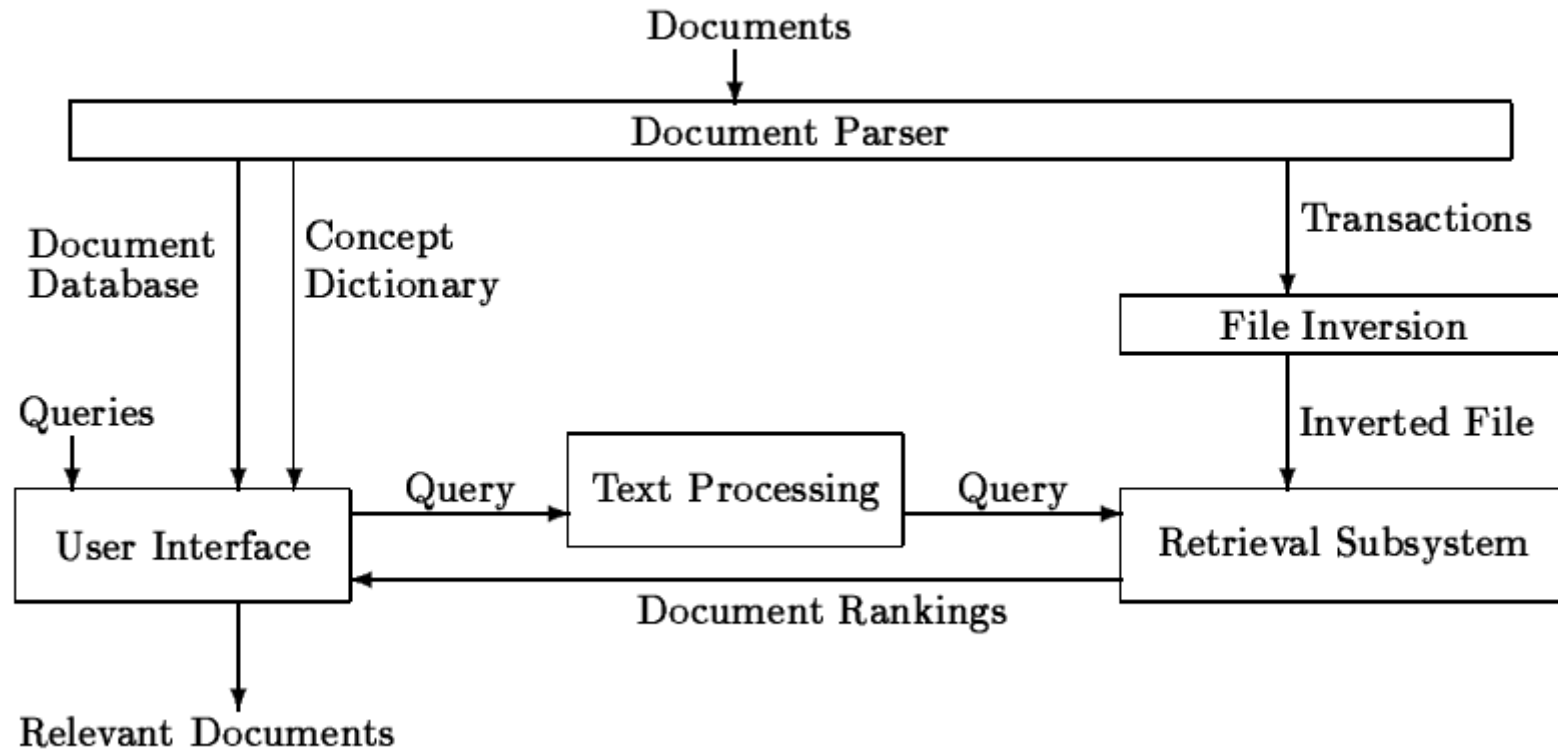
File inversion

- Sort all indexing transaction files
- Build an inverted file for each token

Time distribution to build a document network



INQUERY Architecture



Retrieval subsystem

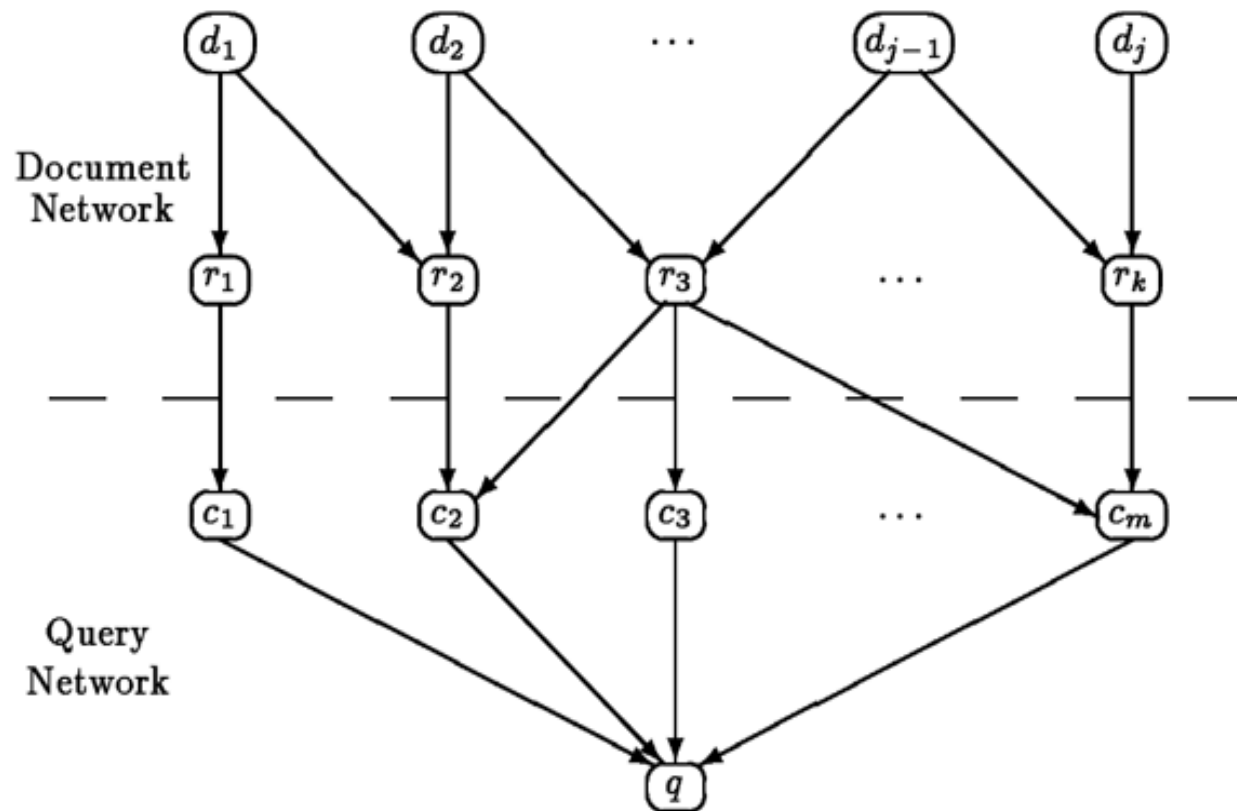
- Build a set of query nodes
- Build proximity lists for each node
- Build a belief list by evaluating all entries in all proximity lists using scoring and weighting function

Query language

OPERATOR	ACTION
#and	AND the terms in the scope of the operator.
#or	OR the terms in the scope of the operator.
#not	NEGATE the term in the scope of the operator.
#sum	Value is the mean of the beliefs in the arguments.
#wsum	Value is the sum of weighted beliefs in the arguments, scaled by the sum of the weights. An additional scale factor may be supplied by the user.
#max	The belief is the maximum of the beliefs in the arguments.
#n	A match occurs whenever all of the arguments are found, in order, with no more than n words separating adjacent arguments. For example, #3 (A B) matches "A B", "A c B" and "A c c B".
#phrase	Value is a function of the beliefs returned by the #3 and #sum operators. The intent is to rely upon full phrase occurrences when they are present, and to rely upon individual words when full phrases are rare or absent.
#syn	The argument terms are to be considered synonymous.

Table 1: The operators in INQUERY's query language.

A simple Bayes network used in INQUERY



Numbers

- Sample test data – heterogeneous information collection of size 1 Gb.
- Transaction files produced by the system – 1.3 Gb's.
- Inverted file size – 0.88 Gb.
- Time to build a document network – roughly 20 hours.
- Time to process a typical query – about a minute.