



Overview

Information Retrieval

Some Current Research in Information Retrieval at CRL

Mark Davis & Bill Ogden

<http://crl.nmsu.edu/Research/Projects/tipster/ursa>



Overview

Information Retrieval

Overview

- MMSUA: Many More Semi-Useless Acronyms.
- Information retrieval and text retrieval
- Design patterns for text retrieval systems
- Evaluating text retrieval systems
- Example: CLTR and QUILT
- Example: Interactive experiments with Infoview



Acronyms...

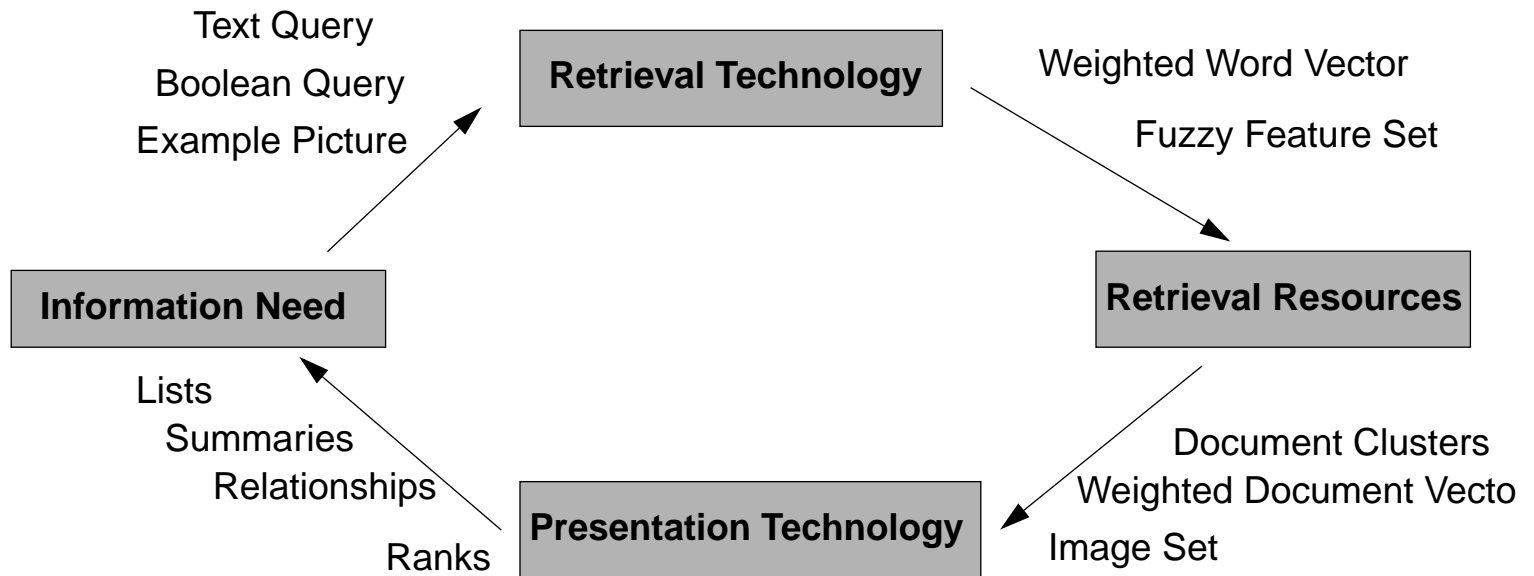
- IR=Information Retrieval
- TR=Text Retrieval
- TREC=Text Retrieval Evaluation Conference OR Text REtrieval Conference
- NIST: National Institute for Standards and Technology
- CLTR=Cross-Language Text Retrieval
- IDF=Inverse Document Frequency



Overview

Information Retrieval

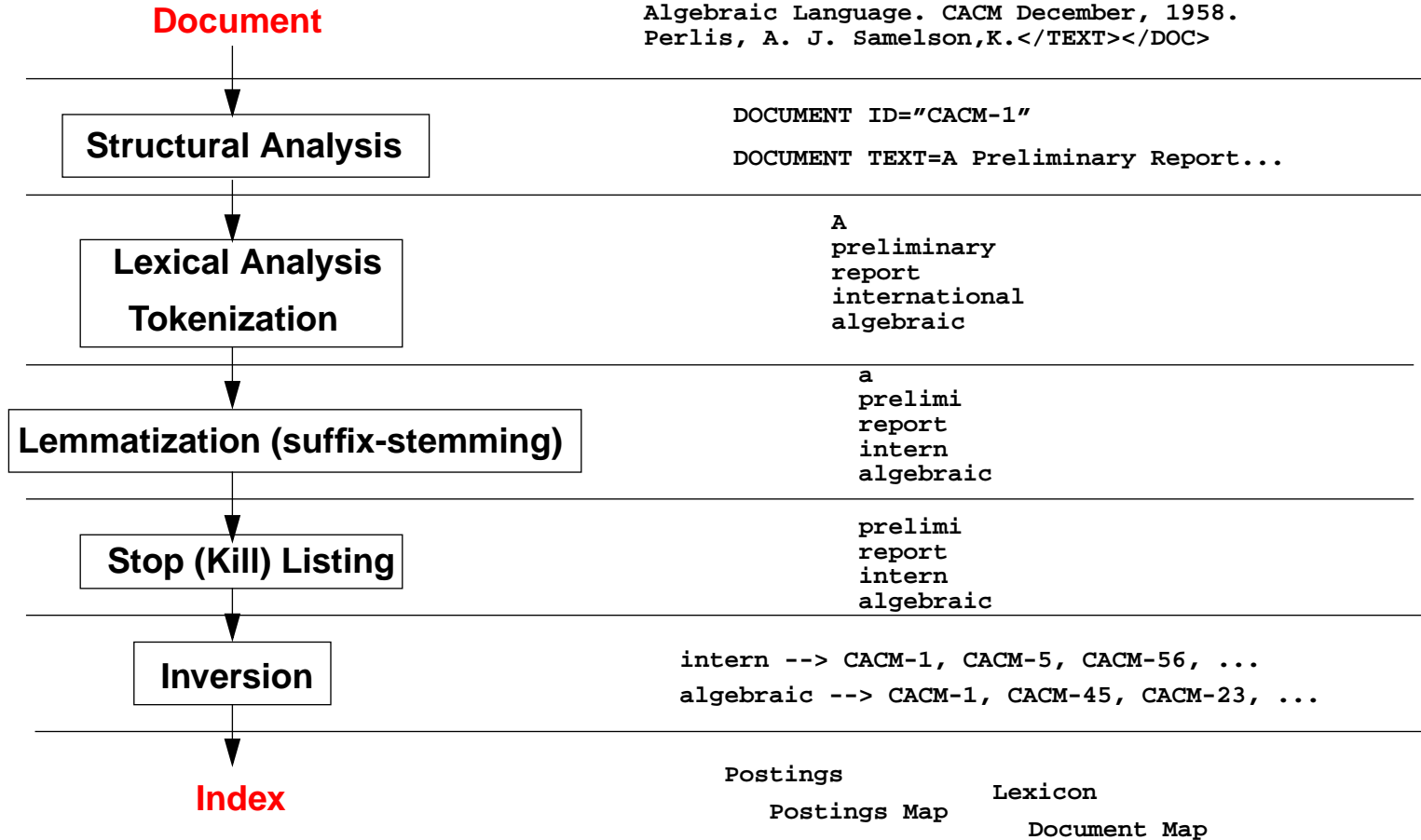
Information Retrieval





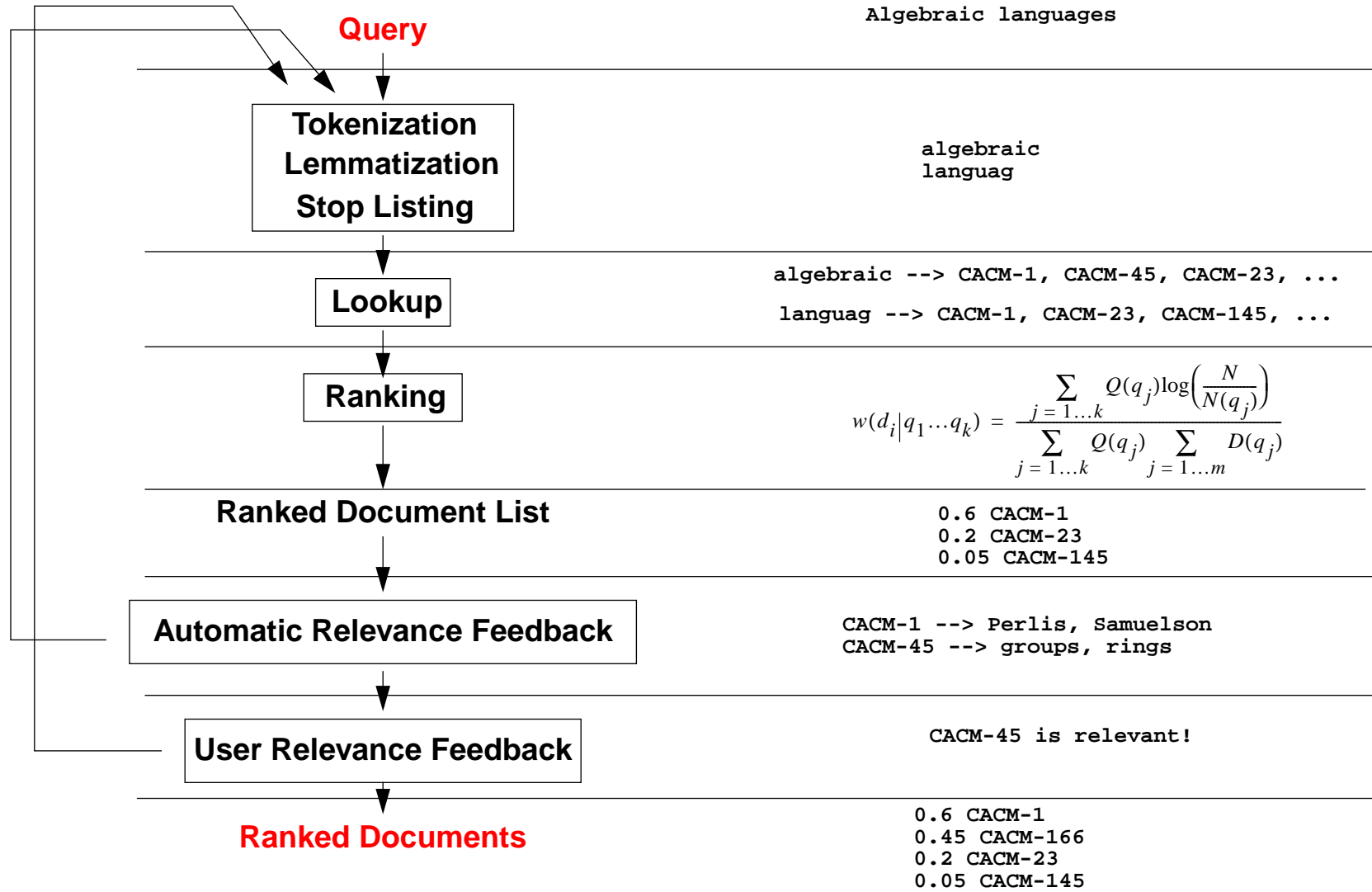
Design Patterns for Text Retrieval: Indexing

```
<DOC> <DOCNO> CACM-1</DOCNO> <TEXT>
A Preliminary Report On An International
Algebraic Language. CACM December, 1958.
Perlis, A. J. Samelson,K.</TEXT></DOC>
```





Design Patterns for Text Retrieval: Querying





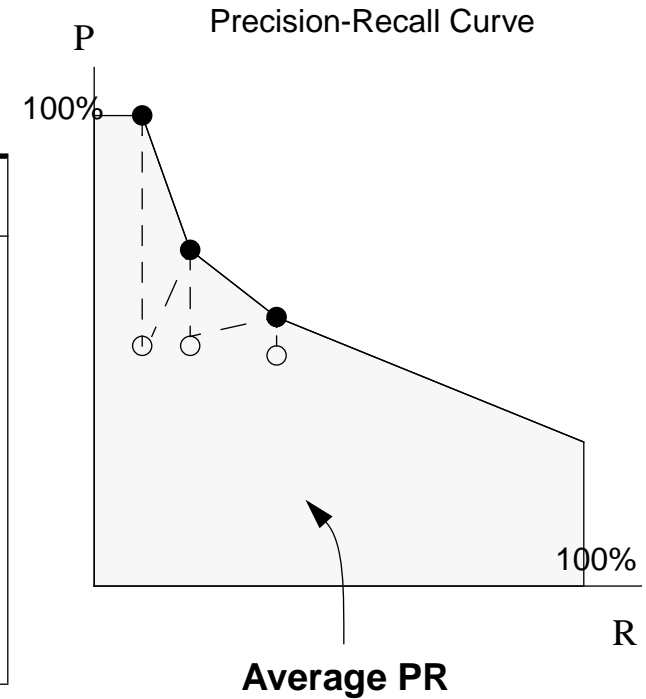
Evaluating Text Retrieval Systems

- Small-scale: experts evaluate the “relevance” of every query to every document
- Large-scale (TREC): experts evaluate the “relevance” of every query to every unique document supplied by the systems (pooling method). Drawback: no easy way to assess how many relevant documents there really are.
- Recall: $R = \frac{\text{\# of relevant documents retrieved}}{\text{\# of relevant documents}}$
- Precision: $P = \frac{\text{\# of relevant documents retrieved}}{\text{\# of documents retrieved}}$



Evaluating Text Retrieval Systems II

Score	Document	Rel?	P	R(10)
0.6	CACM-1	Y	100%	10%
0.3	CACM-45	N	50%	10%
0.2	CACM-23	Y	66%	20%
0.1	CACM-15	N	50%	20%
0.05	CACM-165	Y	60%	30%
0.06	CACM-75	N	50%	30%





QUILT: Query User Interface with Light Translations

- CLTR system: English Queries --> Spanish Documents --> English Gloss Translations
- Prototype
- Syntactic and semantic disambiguation of query terms
- Semantic disambiguation using information retrieval approaches
- Delivered and tested on large-scale collections



Query Translation

- **English Query**

How has the threat of swine fever affected international trade?

- **Tagged English Query**

VB_has NN_threat NN_swine
NN_fever VB_affected
JJ_international NN_trade

- **Stemmed Query**

NN_threat NN_swine NN_fever
VB_affect JJ_intern
NN_trade

- **Dictionary Lookup**

NN_threat achor|amag|amenaz|bravat|conmin|disfuerz|espant|nubla
NN_swine canall|cochin|galduf|jet|malaj|mam|marran|papal|perr|
NN_fever calentur|chuch|fiebr|pasm
NN_intern intern
NN_trade comerc|contrat|negoc|ofic|sindicat|tráfag|tráfic|trapi

- **Corpus Disambiguated Query**

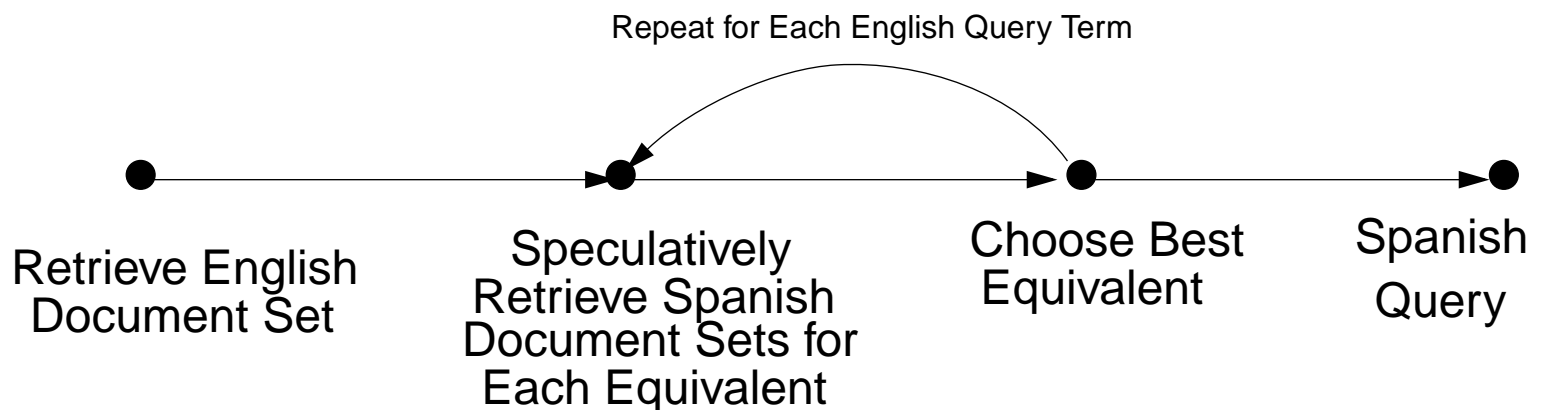
amenaz perr fiebr afect intern comerc

- **Spanish Retrieval**



Corpus Disambiguation

- Parallel English-Spanish aligned corpus (UN 1991).
- Around 100,000 alignment pairs.
- Alignments at the sentence or sentence-pair level.





Overview

Information Retrieval

Performance

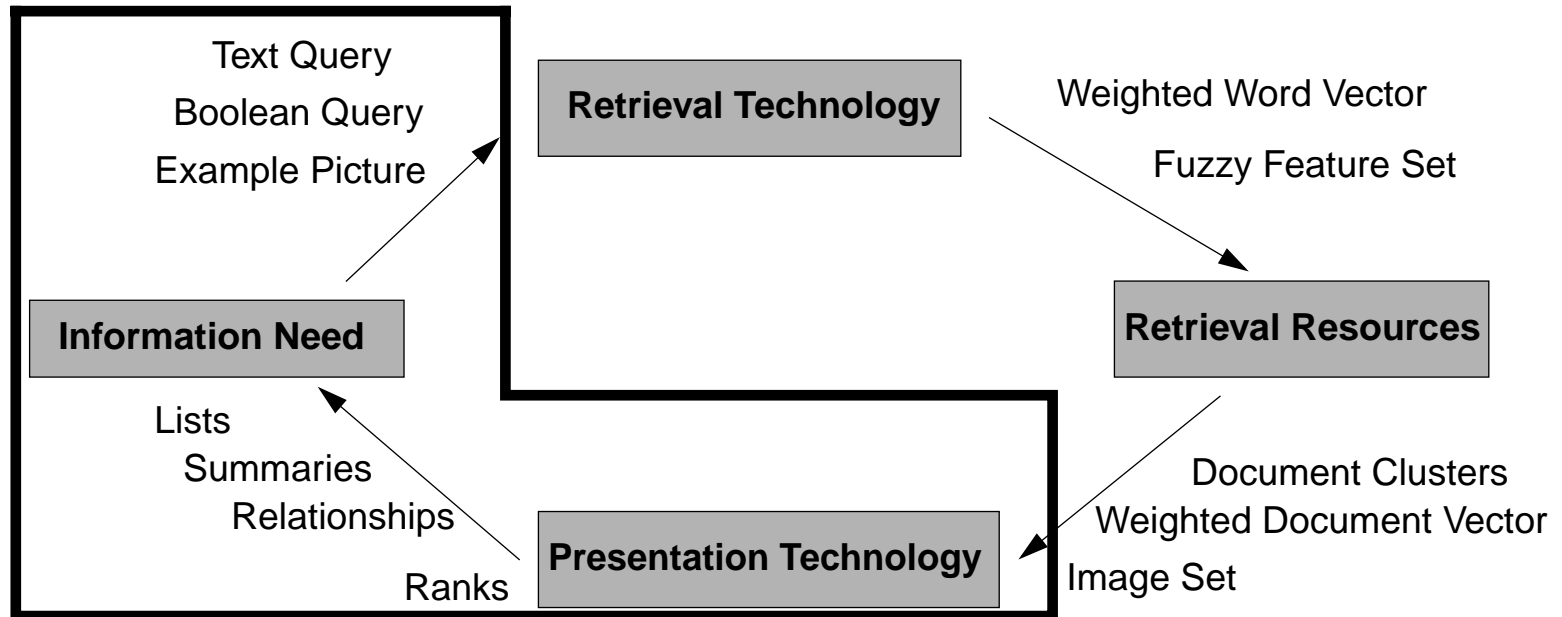
Method	PR (NI)	% of MONO
MONO	0.2895	—
QUILT	0.2127	73.5
POS	0.1949	67.3
ALL	0.1422	49.1
CORP	0.1153	39.8



Overview

Information Retrieval

TREC Interactive Track





Getting better queries from users

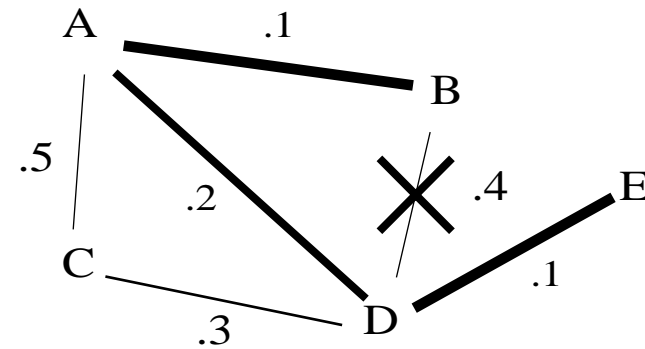
- Users are the best source of information content in queries
- Better queries come from users who “know the data”
- But ad-hoc users, by definition, don’t know the proper terms and relationships represented in the data
- A network representation of databases content is used to assist users in constructing good queries



Pathfinder Networks

$1/f(XY) = 1/\text{frequency of terms together}$

	A	B	C	D	E
A		.1	.5	.2	-
B			.2	.4	-
C				.3	-
D					.1



Shorter 2-link paths beat single links!



Query Modification Using Networks

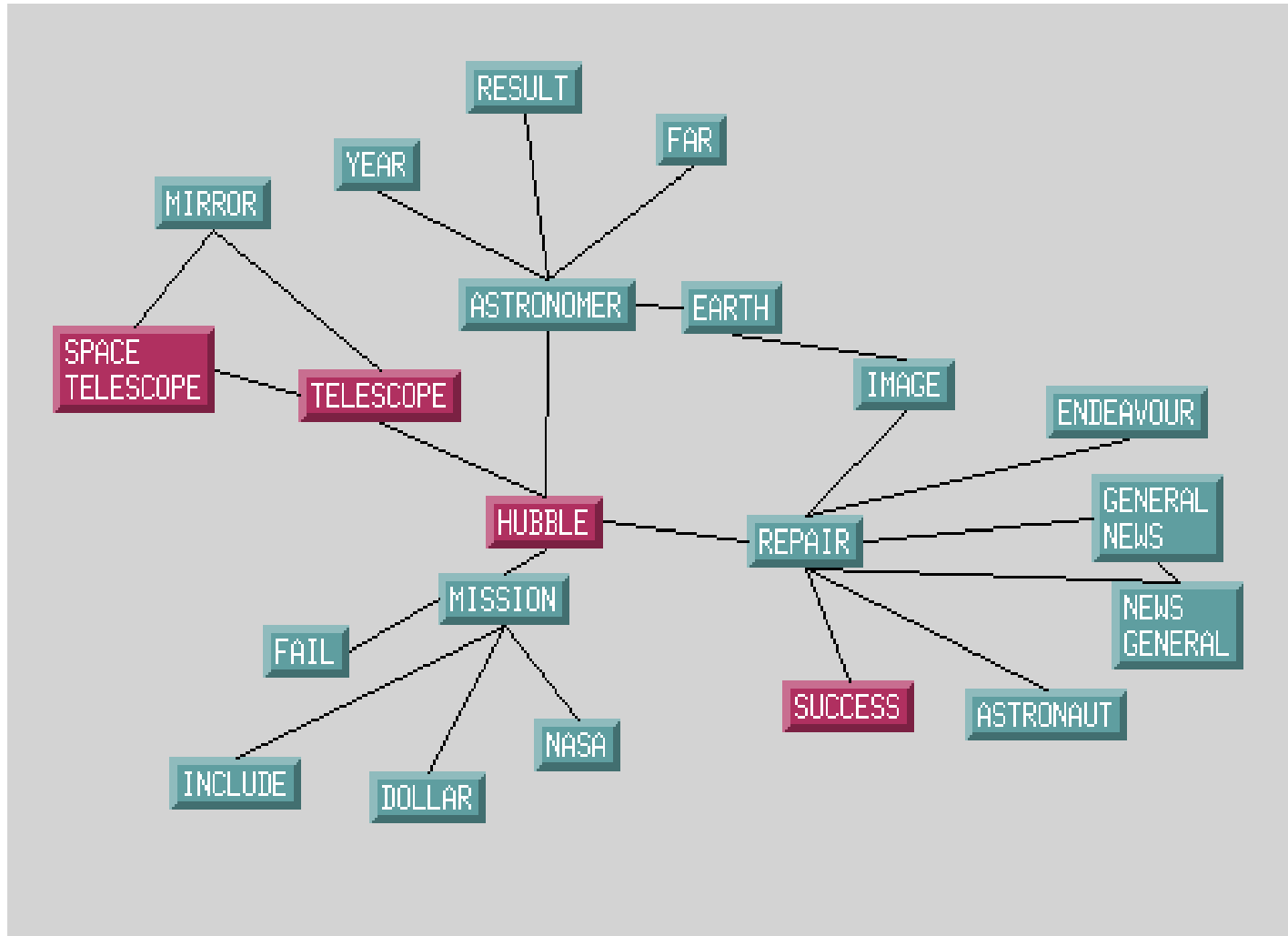
- InfoView provides a network representation of the database to facilitate the process of query modification
- The networks are used to improve the information statements
- Users can substitute precise index terms for vague ones
- Can remove or “negatively weight” inappropriate index terms
- And can add appropriate index terms



Overview

Information Retrieval

Infoview





TREC Interactive Methodology

- Finding different “aspects” of topics
 - Aspectual recall: fraction of all aspects (determined by NIST assessor) covered by saved documents
 - Aspectual precision: fraction of saved documents covering at least one aspect
 - Elapsed time: maximum 20 min.
- Four searchers using Infoview and Zprise (a control system) each for three topics trying to save one document for each aspect.



Aspectual Search Example

<title> Ferry Sinkings

<desc> Description: Any report of a ferry sinking where 100 or more people lost their lives.

<narr> Narrative: To be relevant, a document must identify a ferry that has sunk causing the death of 100 or more humans. It must identify the ferry by name or place where the sinking occurred. Details of the cause of the sinking would be helpful but are not necessary to be relevant. A reference to a ferry sinking without the number of deaths would not be relevant.

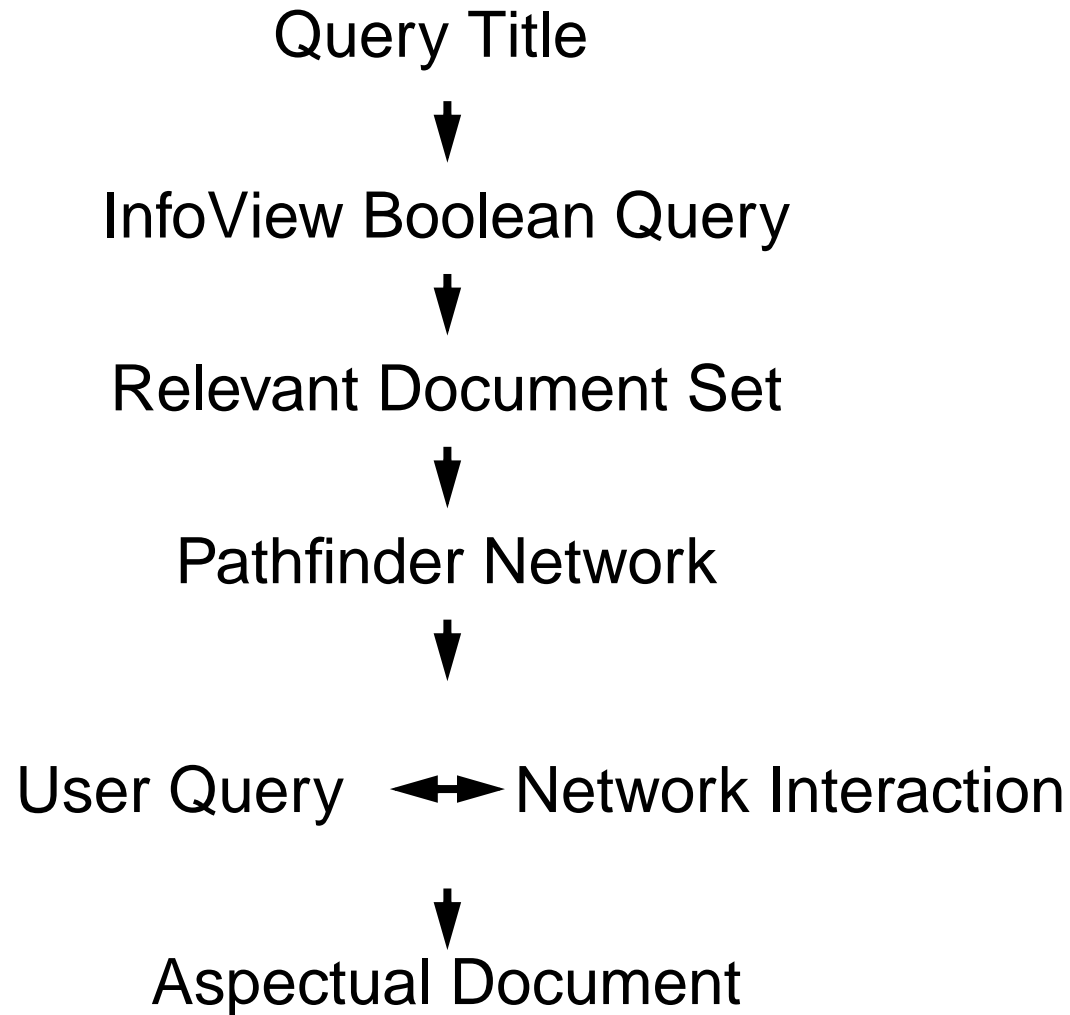
<aspects> Aspects: Please save at least one RELEVANT document that identifies EACH DIFFERENT ferry sinking of the sort described above. If one document discusses several such sinkings, then you need not save other documents that repeat those aspects, since your goal is to identify different sinkings of the sort described above.



Overview

Information Retrieval

Interaction with Infoview





Interactive Summary

- Networks using homogenous data give better results
- Interactive track data suggest InfoView users took less time and were at least as good as ZPRISE users
- Further work with clustering techniques and LSI will be designed to improve the usefulness of the network visualizations
- Investigate “Aspectual Change”
 - Aspects known before interaction / aspects known after
 - Measure the effectiveness of the interaction