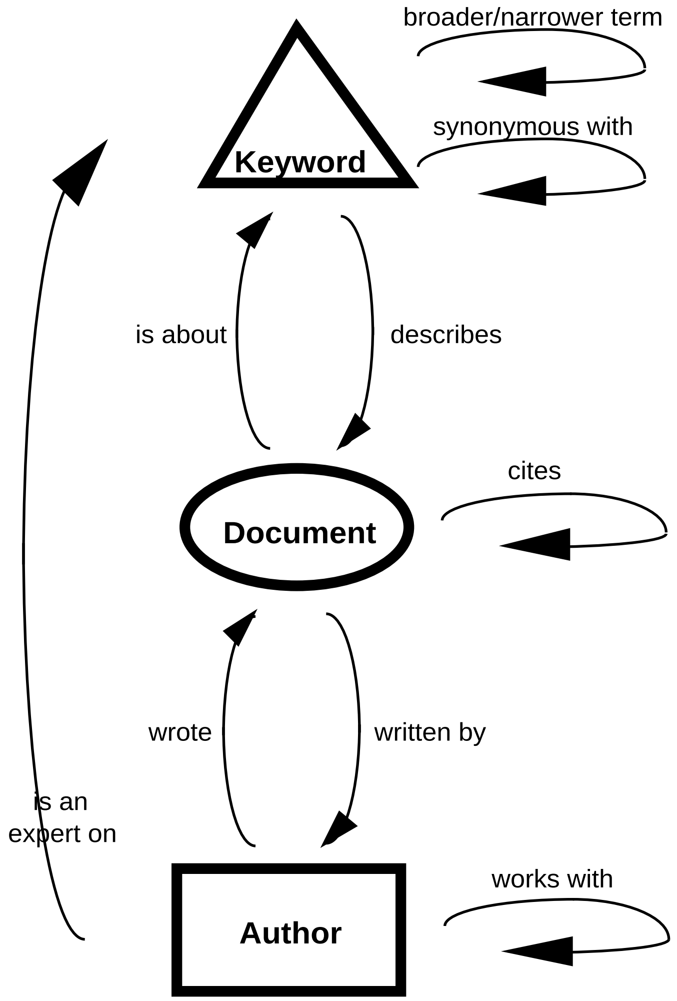
Background

- History: IR xor AI
- Using manual indices as benchmark, training set
- Heterogeneous data sources
- Info access especially ripe for machine learning

History: IR xor AI

- In 1960's, "deep understanding" of text promised by AI/NLP methods made IR's statistical character "shallow"
- Now: Both machine learning and corpus-based linguistics share very similar statistical methods with IR

Beyond indexing (keyword/document) information



- Database-style attributes can be useful
- Learning implies use of syntactic “clues” only
- Ie, no knowledge-intensive coding
    - At least for every document

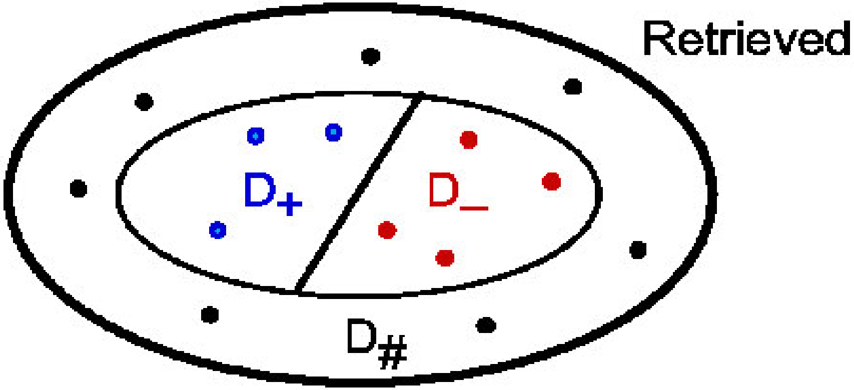
Info access especially ripe for machine learning

- Always more readings than writings
  - | Queries | >> | Documents |
- Relevance feedback
  - Assessments by users of retrieved documents
  - Used for changing qry
  - Used for changing docs

Info access especially ripe for machine learning (cont.)

- IR appeals to both symbolic (logical) and statistical ML
  - Documents' words are intrinsically "meaningful"
  - But can also be treated as "meaningless" features

Co-relevance relation



- Assessments are context sensitive
  - $D_{++} > D_{+} > D_{\#} > D_{-} > D_{--}$
- How to use negative feedback?
  - D_+ plausibly clustered; lots of reasons for D_-

Learning from browsing across queries in same "session"

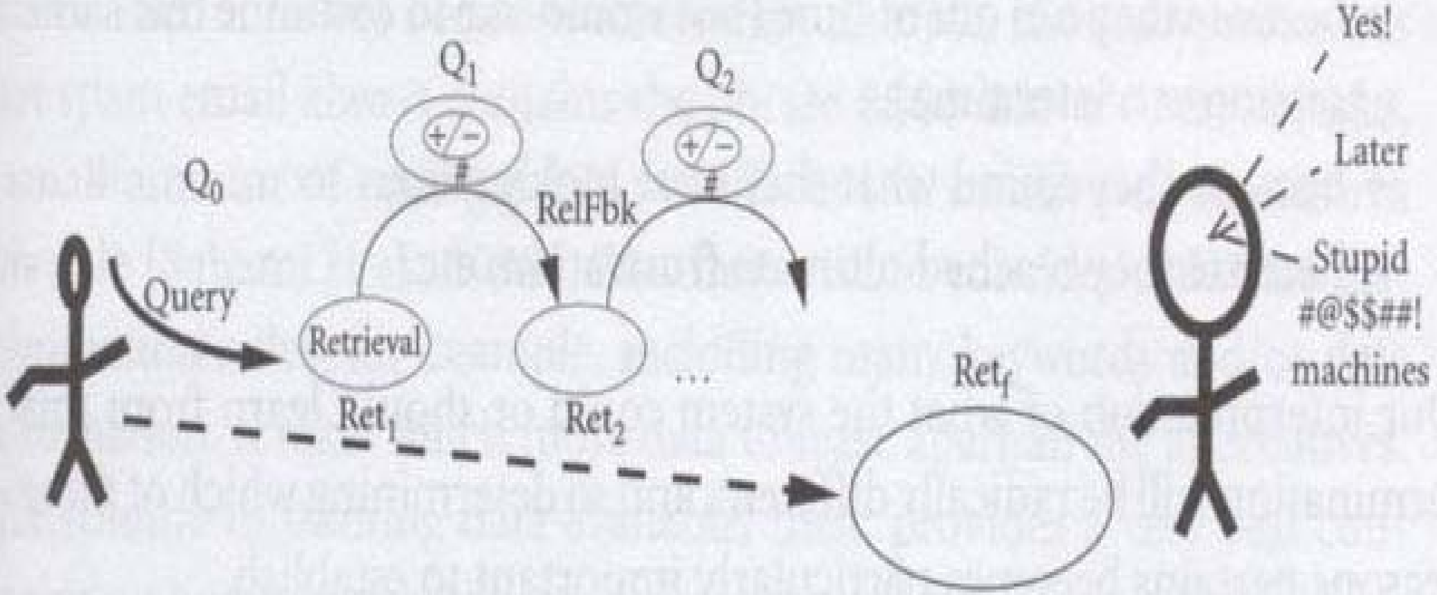
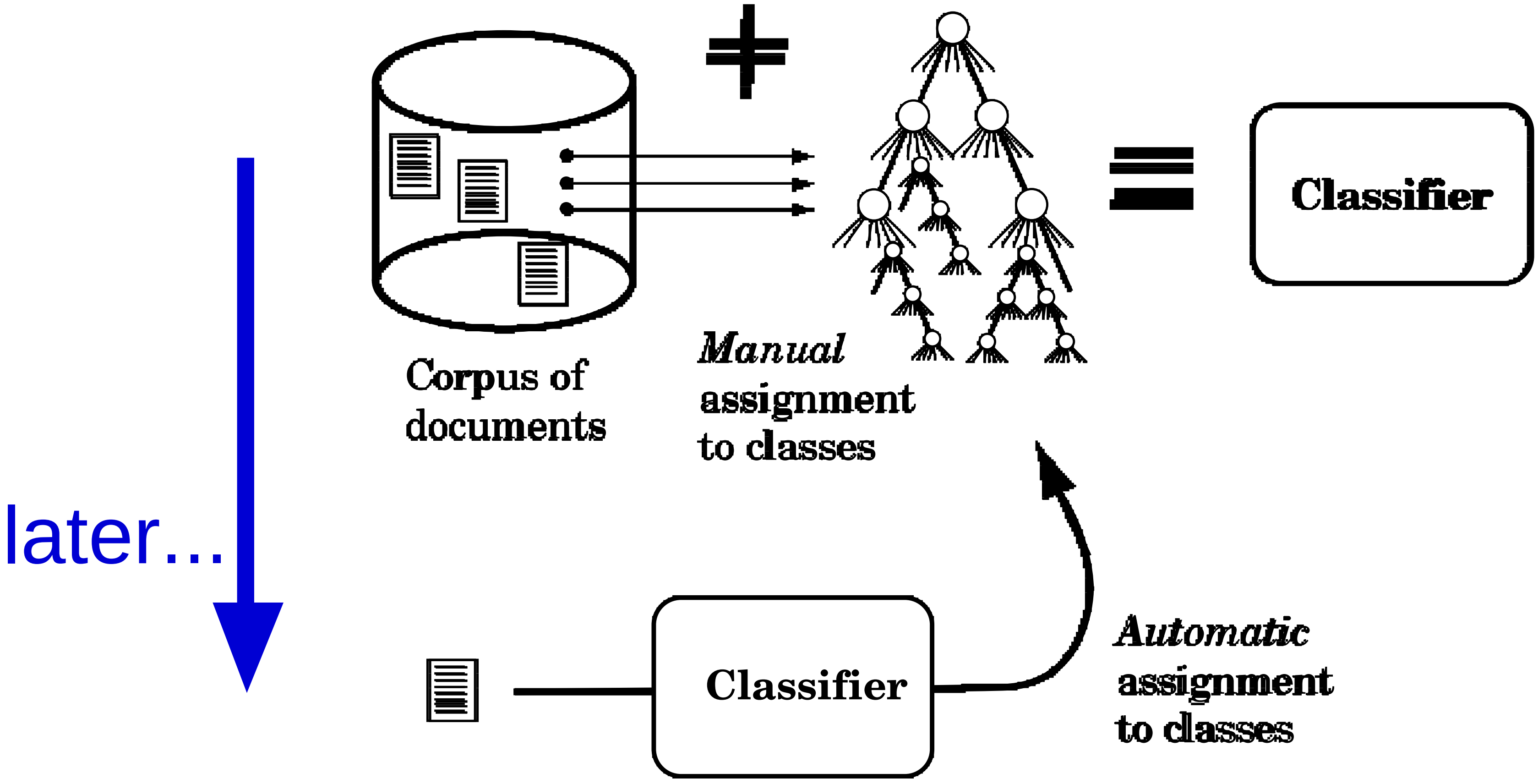


FIGURE 7.2 Browsing across Queries in Same Session

Training a classifier



Corpus of documents

Manual assignment to classes

Classifier

later...

Classifier

Automatic assignment to classes

Characteristics of training set

- #features/#instances ration
- /- instance ratio

Attribute/Feature vectors

- Feature selection
- “Hypothesis spaces” formed over features
- “Inductive bias”

Feature selection

- Obvious features = keywords
- Less obvious features
- Feature clustering

Obvious features = keywords

- Documents characterized by large (sparse) vectors!
  - => “irrelevant attributes abound” [Littlestone]
- Mutual information
  - [vanR]
  - [Yang, Pedersen, ICML97]

Less obvious features

- “Singulars” (proper names)
- Biblographic citations

“Hypothesis spaces” formed over features

- Boolean functions
    - even k-DNF [Moulinier’96]
        - inductive bias = disj. of conj. each fitting small sets of examples
- Linear combinations
- Non-linear functions
- Bayesian networks!

“Inductive bias”

- Parsimony: Occam’s razor
- Over-fitting available training data

Learning to improve retrieval effectiveness

- Intro
- Prediction
  (vs Classification, ...)
- Document(s) vector modification!

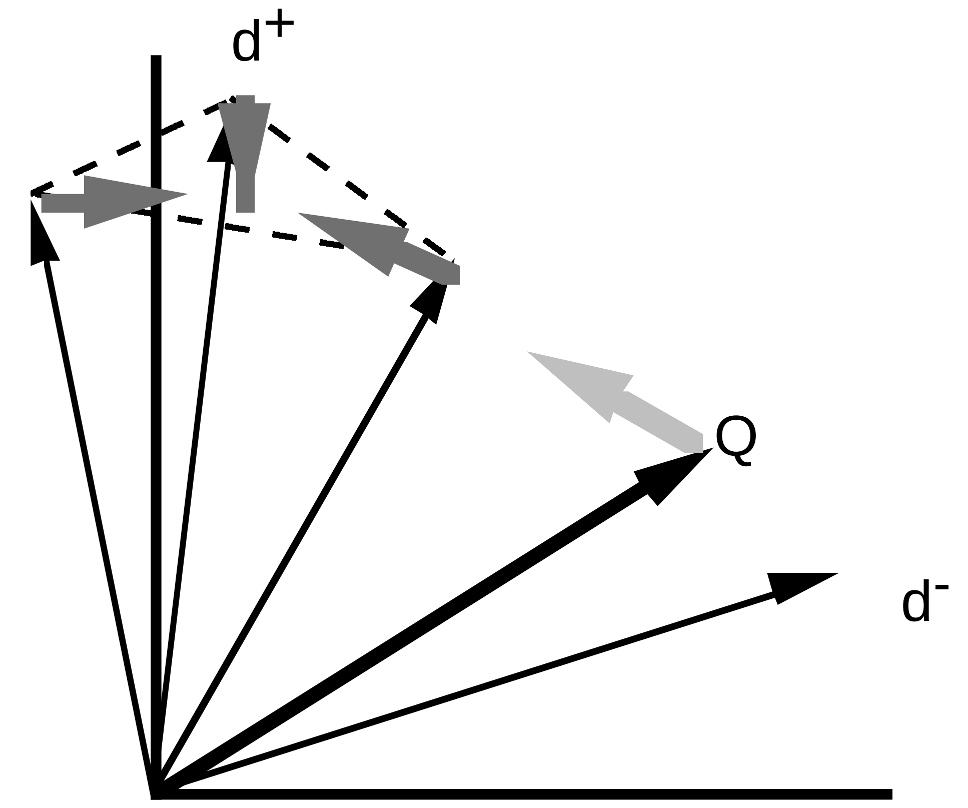
Intro

- ala [Roccio]
- Goal: Improve retrieval effectiveness (precision/recall)

Prediction
(vs Classification, ...)

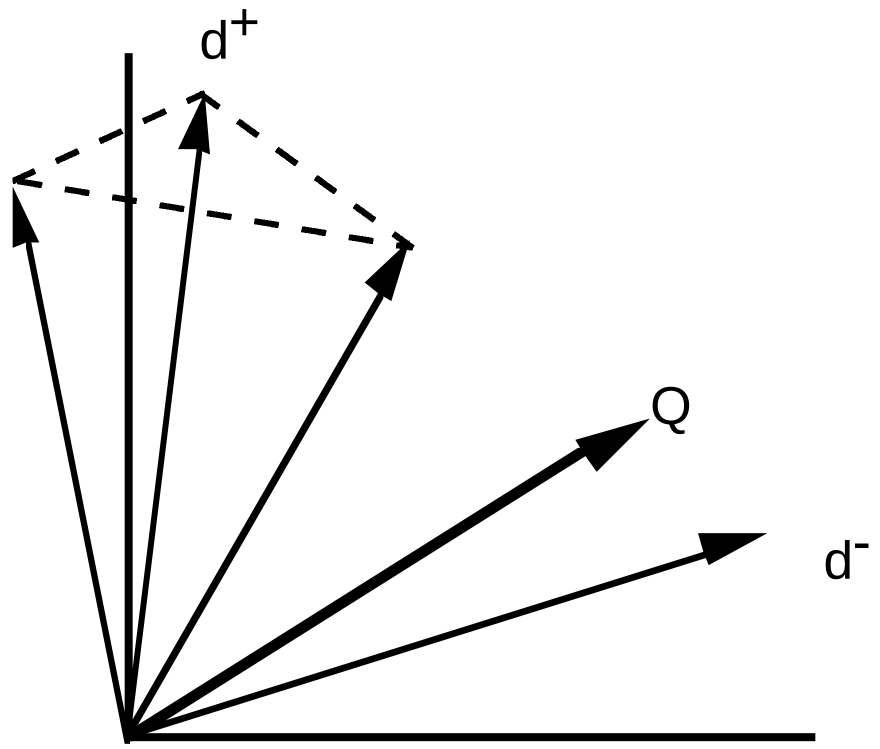
- f(attributes) -> \Real
- Pr(Relevance), next set to be ‘classified’

Document(s) vector modification!



- Recall: Query vector modification

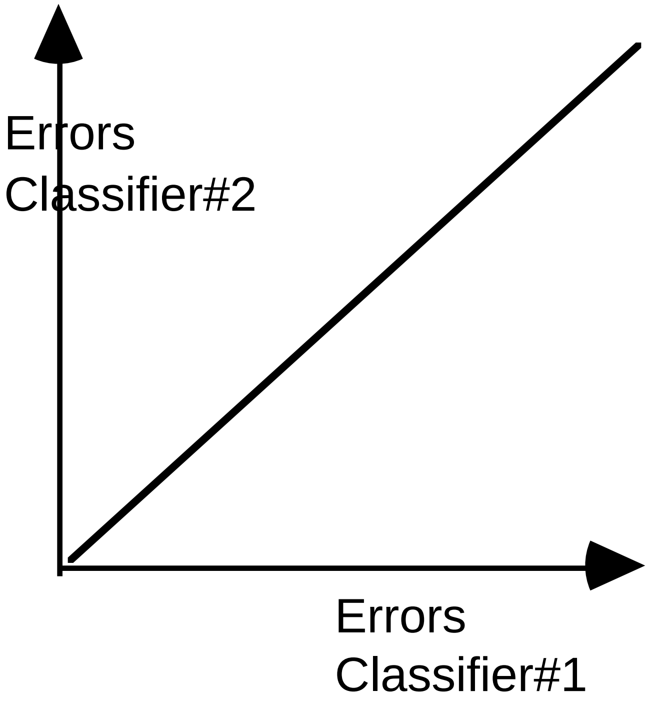
Recall: Query vector modification



- Corresponds to first-order associates

Classification

- Assigning documents to classif
- Naive Bayes - the Canonical classifier
- Evaluating classifiers
- Other learning algorithms
- Combining classifiers
- Hierarchic classification



Assigning documents to classif

- f(attributes) -> Class_i
- Binary (ir/relevant) classes
- Also routing/filtering
- Assignments

Assignments

- Binary/Boolean
- Multi-class: Pr(class)

Naive Bayes - the Canonical classifier

- Assume parametric model
- Connections to IR Probability
- [Lang, Mitchell]

Assume parametric model

- (used to generate docs)
- Training data used to estimate parameters

Connections to IR Probability

- Priors:
- Density vs. condidtional density (regression) [Lewis,51]
- Two underlying models of “event space”

Priors:

- What we begin knowing
- Term descrimination: moderate frequencies best
- Correlations x_i, x_j

Term descrimination: moderate frequencies best

- Pr(hi) = noise
- Pr(lo) = not much data!

Two underlying models of “event space”

- Multi-variate Bernoulli
- Multinomial
- Entropy distributions (used for MI feature selection) depends
- Normalization on document lengths
- NB: Can't just add non-textual features as part of same urn

Multi-variate Bernoulli

- Biased coin for each term
- Pr(w | class) = #docs in with word W_t/ #docs
- Non-occurring words matter

Multinomial

- Colored balls in urn w/ replacement
- \sum_t Pr(word_t | class) = 1
- Evidence only from words which DO occur; consider their frequency
- Better!! Esp, w/ large vocabularies

Evaluating classifiers

- Train vs. test
- Cross-validation
- F-measure [Lewis]
- Scatter plot of relative classif. errors

Other learning algorithms

- Winnow/EG/Sleeping experts!!!
  - [Allan SIGIR98]: Target = {min,max}
- RIPPER - a logical classifier
  - Covering algorithms (while + not covered...)
- Widrow-Huff
- Decision-tree

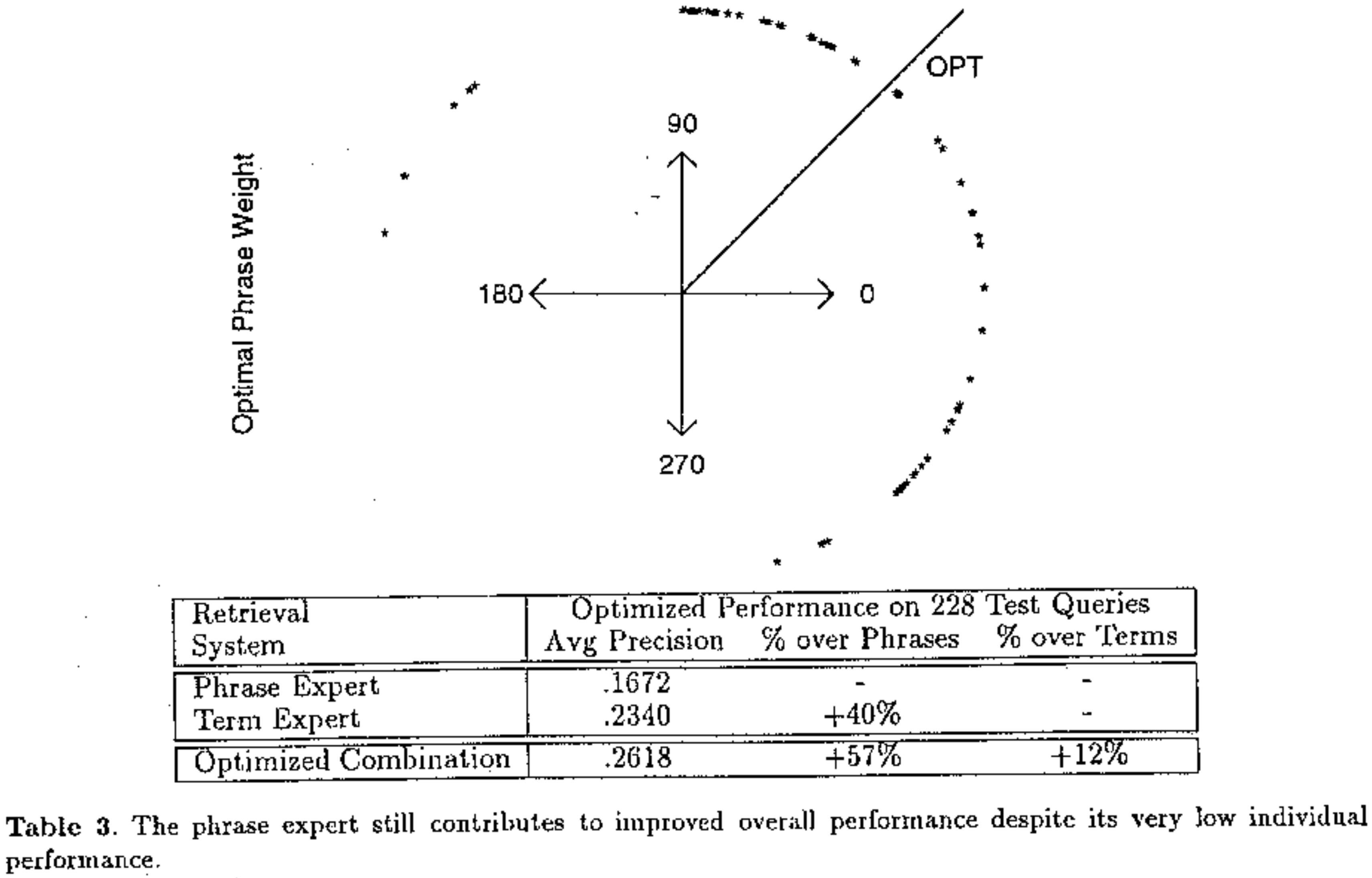
Other learning algorithms (cont.)

- k-Nearest Neighbor
- Other clustering methods?
- Structural Risk Min/Support vectors

Applications/Examples

- Optimal expert weights
- Parameter optimization
- Simpler, more learnable problems
- Retrieval from heterogeneous corpora
- Info-seeking agents

Optimal expert weights



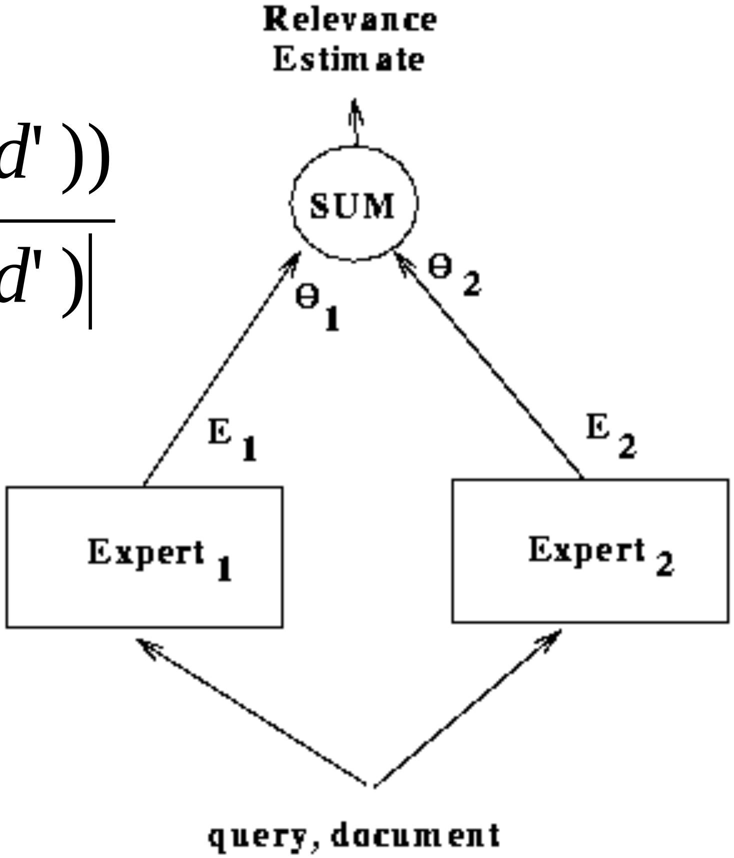
Retrieval System	Optimized Performance on 228 Test Queries Avg Precision	 % over Phrases	 % over Terms
Phrase Expert	.1672	-	-
Term Expert	.2340	+40%	-
Optimized Combination	.2618	+57%	+12%

Table 3. The phrase expert still contributes to improved overall performance despite its very low individual performance.

Parameter optimization

- Point alienation measure on retrieval method **R_θ**

$$J(R_\theta) = \frac{\sum_{d>d'} (R_\theta(d) - R_\theta(d'))}{\sum_{d>d'} \left| R_\theta(d) - R_\theta(d') \right|}$$



Simpler, more learnable problems

- “Learning” corpus statistics
  - Collection KW freq
  - Avg. doc length
- Short queries
- Small corpora
- Query types
- Individual user's preferences
- Web pages that have changed

Retrieval from heterogeneous corpora

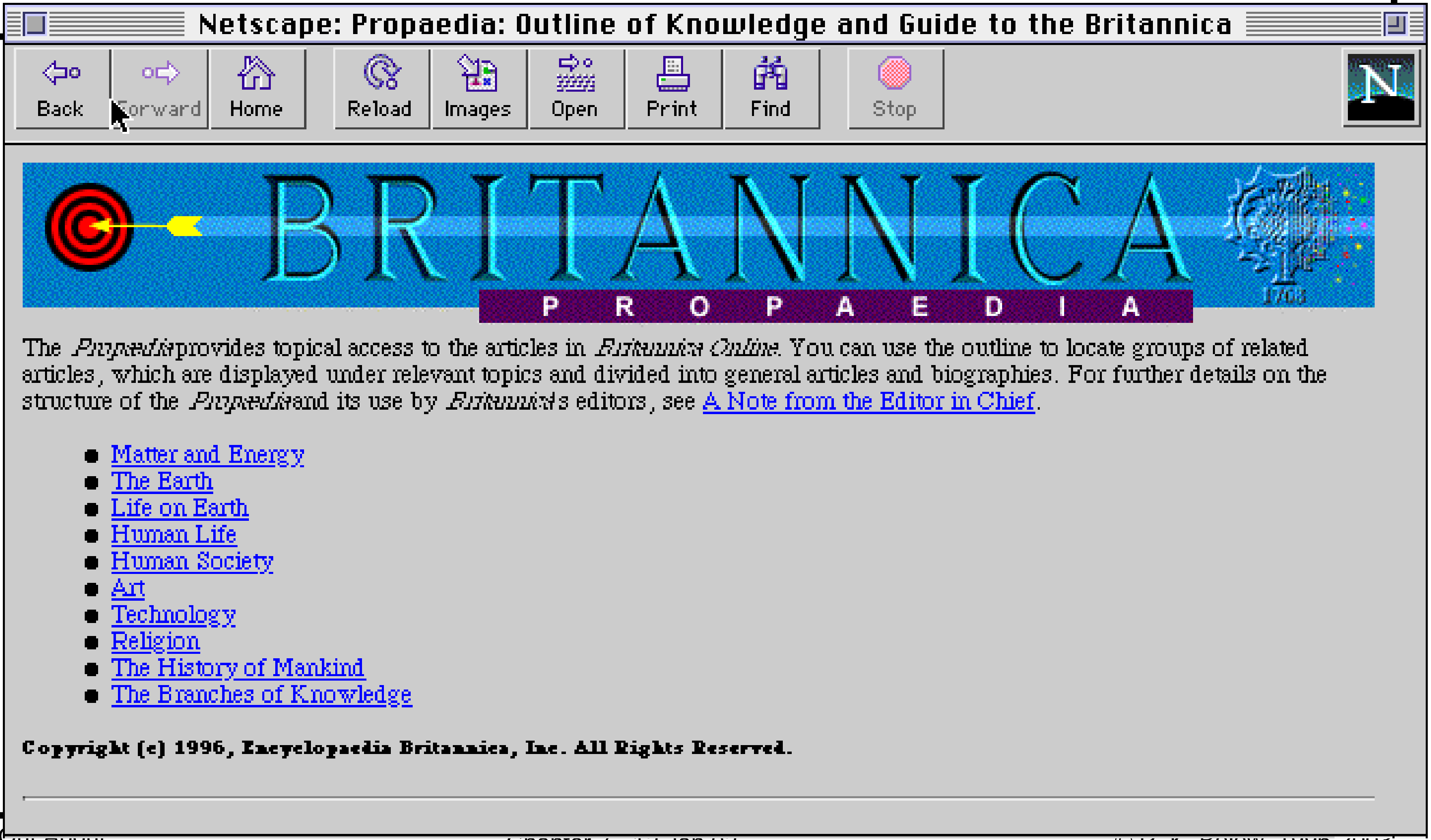
- Data fusion
- Vocabulary mismatch

Info-seeking agents

- Encyclopedia Britannica data
- Basic mapping
- ARACHNID Algorithm
- Final population

Ency. Britannica: CD, Internet versions

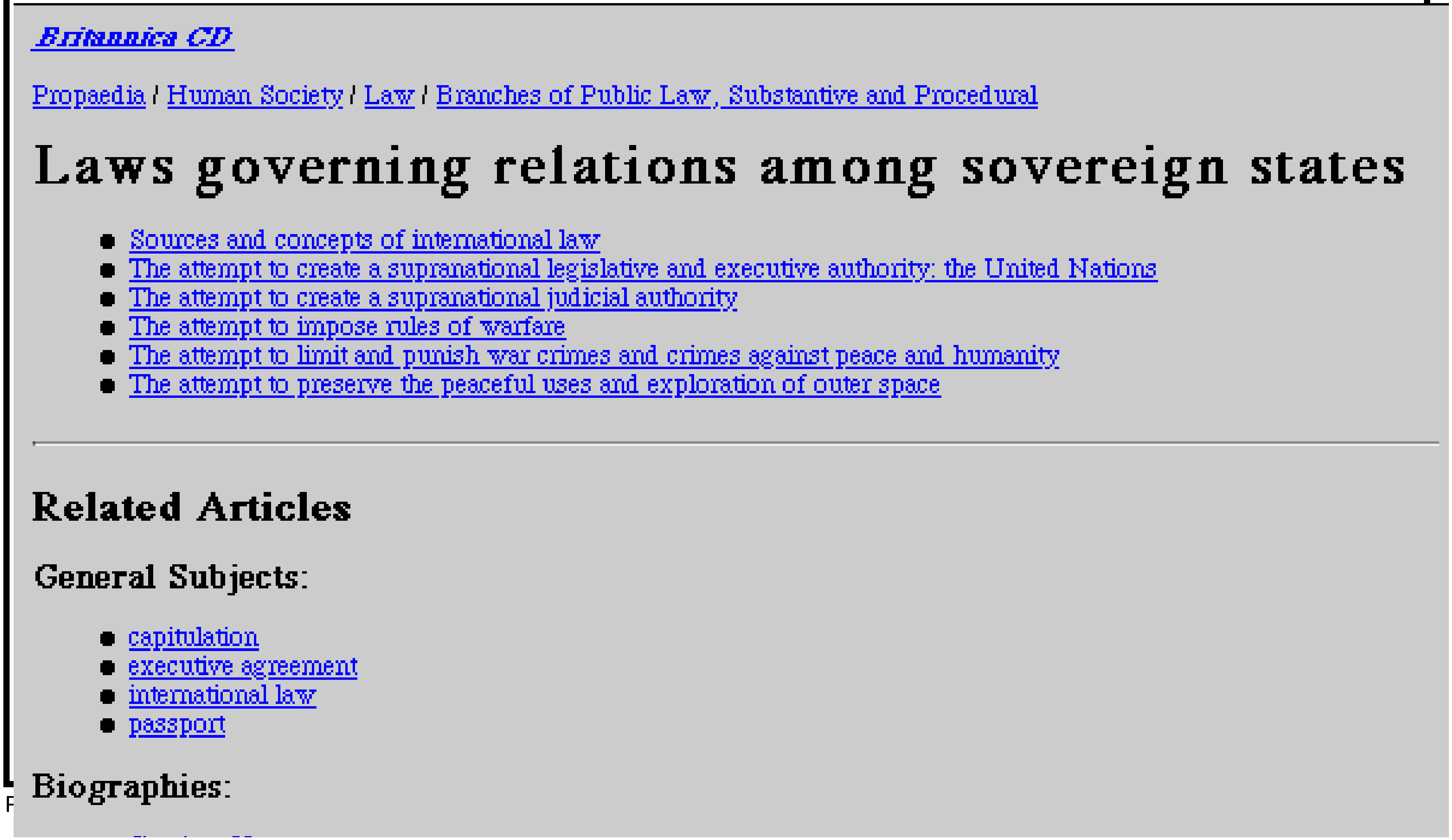
- Propaedia, Micro, Macro


Netscape: Propaedia: Outline of Knowledge and Guide to the Britannica
Back
Forward
Home
Reload
Images
Open
Print
Find
Stop
N
BRITANNICA
PROPAEDIA
1768

The *Propaedia* provides topical access to the articles in *Britannica Online*. You can use the outline to locate groups of related articles, which are displayed under relevant topics and divided into general articles and biographies. For further details on the structure of the *Propaedia* and its use by *Britannica*'s editors, see A Note from the Editor in Chief.

- Matter and Energy
- The Earth
- Life on Earth
- Human Life
- Human Society
- Art
- Technology
- Religion
- The History of Mankind
- The Branches of Knowledge

Query=Intermediate Propaedia rubric



Britannica CD

Propaedia / Human Society / Law / Branches of Public Law, Substantive and Procedural

Laws governing relations among sovereign states

- Sources and concepts of international law
- The attempt to create a supranational legislative and executive authority: the United Nations
- The attempt to create a supranational judicial authority
- The attempt to impose rules of warfare
- The attempt to limit and punish war crimes and crimes against peace and humanity
- The attempt to preserve the peaceful uses and exploration of outer space

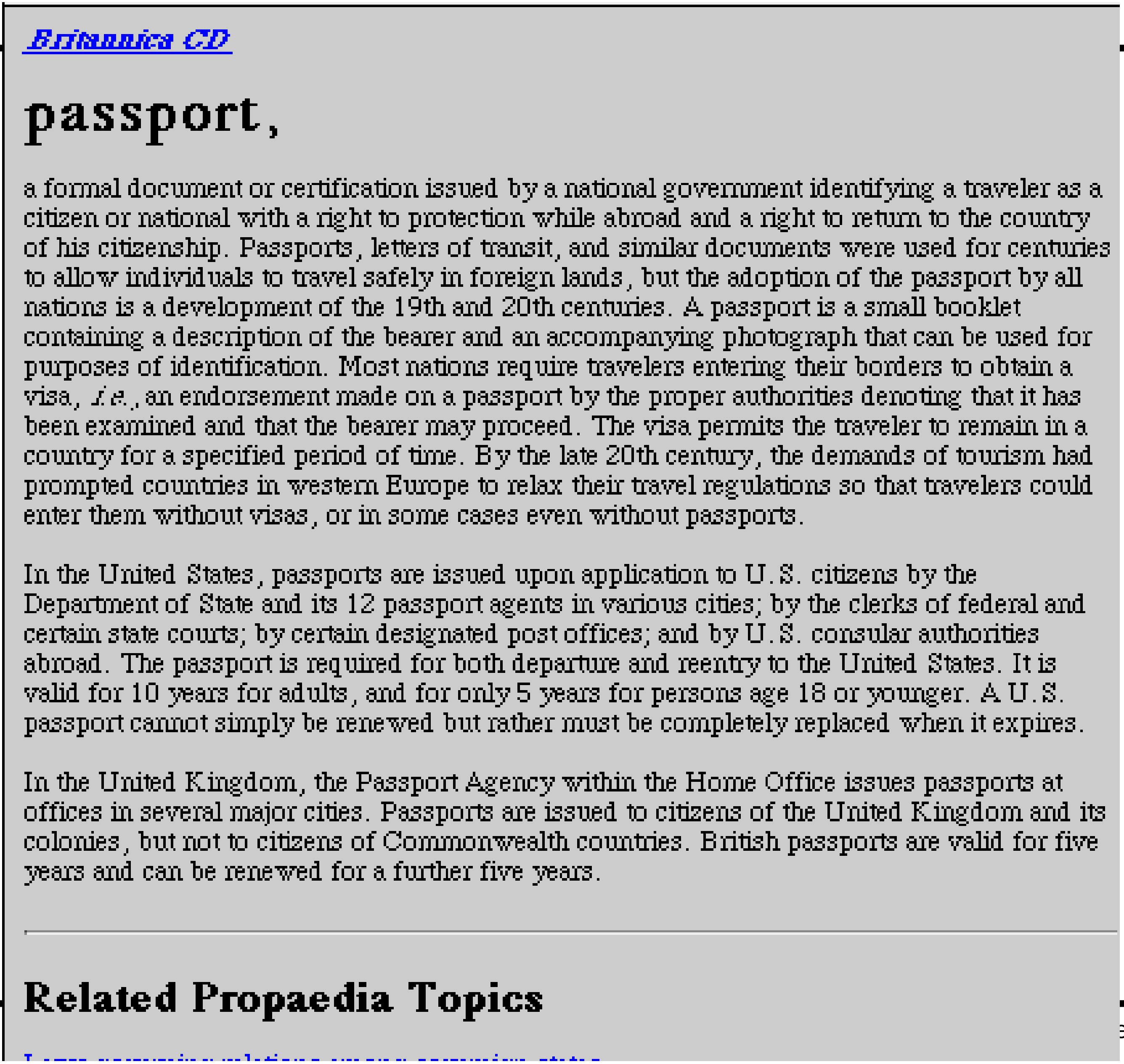
Related Articles

General Subjects:

- capitulation
- executive agreement
- international law
- passport

Biographies:

Typical (Micro) article



Britannica CD

passport,

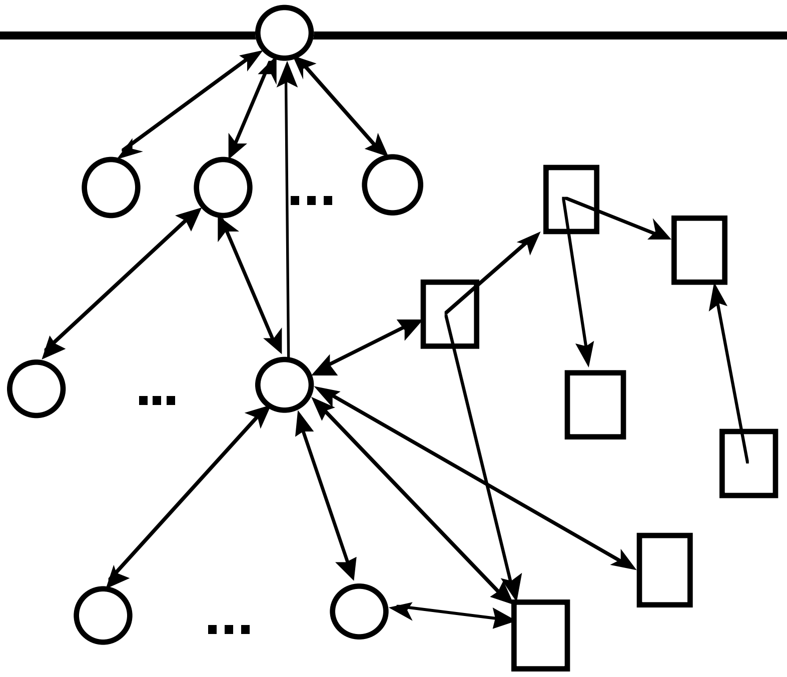
a formal document or certification issued by a national government identifying a traveler as a citizen or national with a right to protection while abroad and a right to return to the country of his citizenship. Passports, letters of transit, and similar documents were used for centuries to allow individuals to travel safely in foreign lands, but the adoption of the passport by all nations is a development of the 19th and 20th centuries. A passport is a small booklet containing a description of the bearer and an accompanying photograph that can be used for purposes of identification. Most nations require travelers entering their borders to obtain a visa, *i.e.*, an endorsement made on a passport by the proper authorities denoting that it has been examined and that the bearer may proceed. The visa permits the traveler to remain in a country for a specified period of time. By the late 20th century, the demands of tourism had prompted countries in western Europe to relax their travel regulations so that travelers could enter them without visas, or in some cases even without passports.

In the United States, passports are issued upon application to U.S. citizens by the Department of State and its 12 passport agents in various cities; by the clerks of federal and certain state courts; by certain designated post offices; and by U.S. consular authorities abroad. The passport is required for both departure and reentry to the United States. It is valid for 10 years for adults, and for only 5 years for persons age 18 or younger. A U.S. passport cannot simply be renewed but rather must be completely replaced when it expires.

In the United Kingdom, the Passport Agency within the Home Office issues passports at offices in several major cities. Passports are issued to citizens of the United Kingdom and its colonies, but not to citizens of Commonwealth countries. British passports are valid for five years and can be renewed for a further five years.

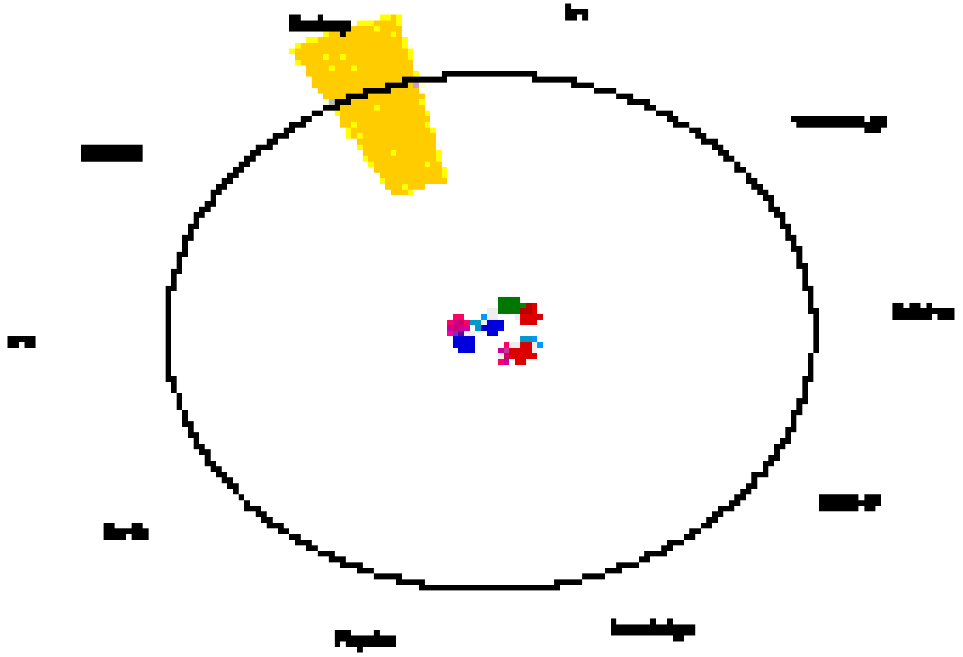
Related Propaedia Topics

EB as graph



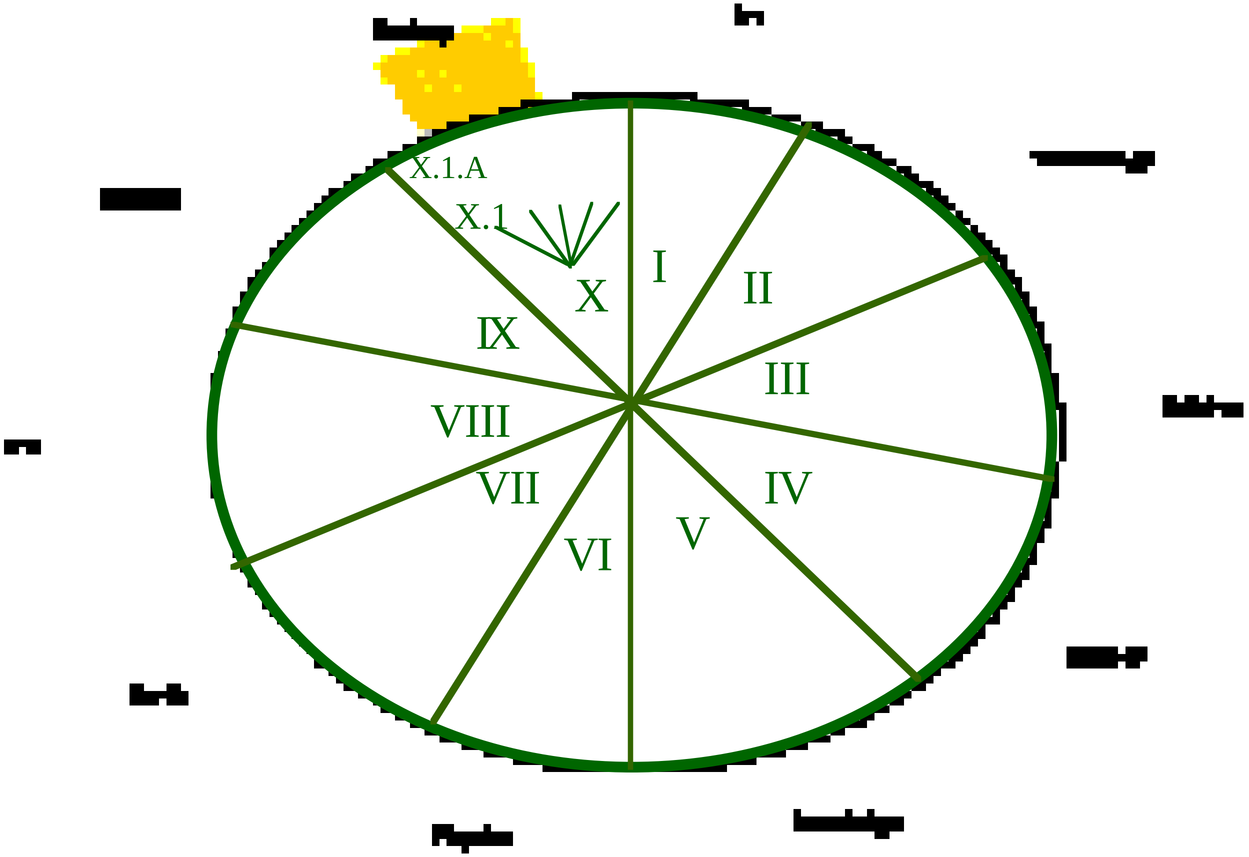
- Propaedia tree ~10 max depth
- Micropaedia articles: (~10^5 articles)
- EB data
- 3 Queries: easy, bad, normal

The EB as “environment”

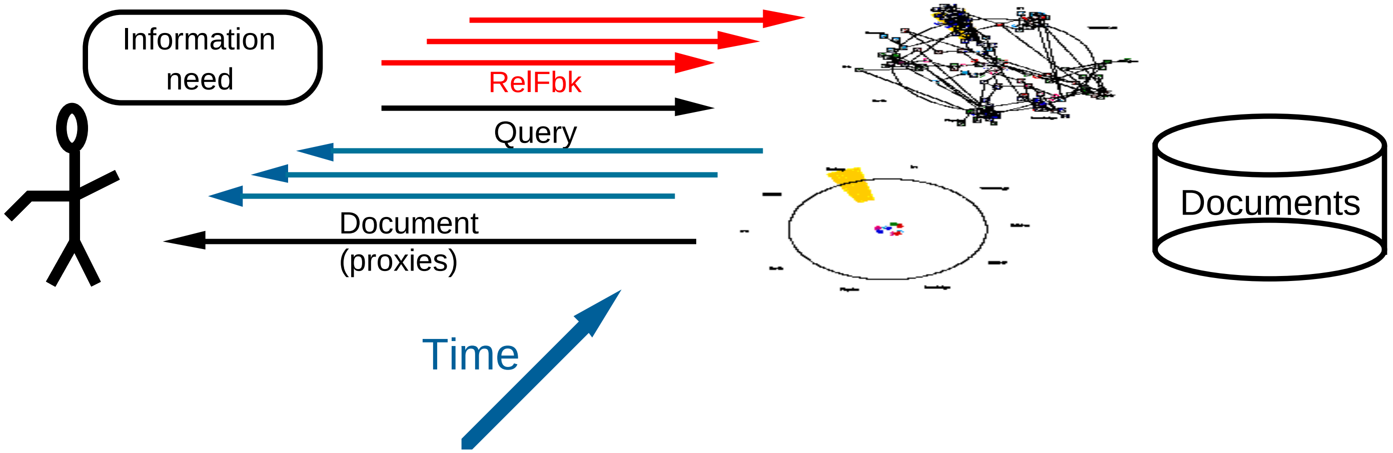


- Crawling over terrain defined by the EB’s links

Propaedia Space



Basic mapping

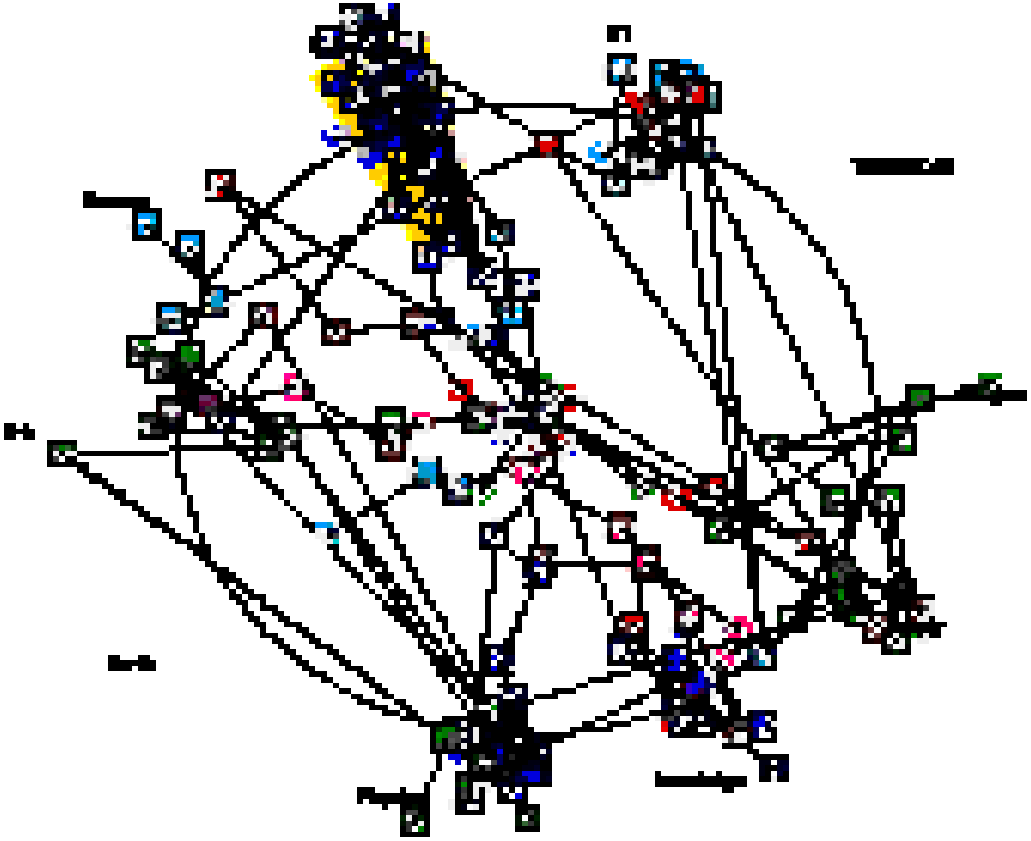


- User RelFbk = ultimate resource
- Agents evolve to efficiently process this resource
- Total population captures shared traits
- Individuals can adapt to ‘local’ characteristics

ARACHNID Algorithm

```
Situate, initialize agents with E = e/2
WHILE there are alive agents:
  pick random agent a
  pick link for a to follow
  fetch new document $D_a$
  $E_a \leftarrow E_a - cost + \begin{cases} r(D_a) & \text{if } D_a \text{ new} \\ f(D_a) & \text{otherwise} \end{cases}$
  mark $D_a$ as visited
  learn by reinforcement
  IF ($E_a$ > e)
    a’ <- mutate(clone(a))
    $E_a, E_{a'} \leftarrow E_a / 2$
  ELSEIF ($E_a$ < 0)
    die(a)
```

Final population



Cultural learning: from others' experience

- Ability to transfer from your experience to mine
- Only statistical reliability required
- Novice vs. expert opinions?
  - Not all opinions equal
  - Experts well-informed, but novices are typical consumers

Cultural learning: from others' experience (cont.)

- Historical time constant
  - How quickly should we adapt to new (current/trendy) vocabulary?
  - How can we maintain archival record of previous terminology?

Learning linguistic problems

- Summarization
- Text segmentation
- Word-sense disambiguation
- Cross-language LSI
  - Corpus graph!