Indexing and Retrieving Structured Documents

Yves Chiaramella
CLIPS-IMAG
BP 43
38041 Grenoble Cedex
France
chiara@imag.fr - http://www-clips.imag.fr/

Introduction

❖ The context

- Fast development of *tools* and *standards* for managing structured documents
  - **Database systems,**
  - **New standards (SGML, XML, HTML, ODA, MPEG …)**
- Increasing user needs about *more focused* system answers
  - **Related to increasing *volume* and *structural complexity* of available documents (Web, hypermedia applications, DB applications)**
  - **Related to increasing diversity of *retrieval situations* (user typology, retrieval context, document types etc.)**

Introduction

❖ Standard IR

In standard IR, documents are considered as *atomic* information units whatever their type or size
As a consequence they are:

• Indexed as a whole

➡ Indexes do not express the internal organisation of the discourse set by the author(s)

• Retrieved as a whole

➡ Users cannot retrieve independant components of documents that might be more adapted (more focused) to their information needs

➡ Questions :

• What is the impact of « structure » on information retrieval?

• If any, then what kinds of approaches to improve retrieval performances?

Plan

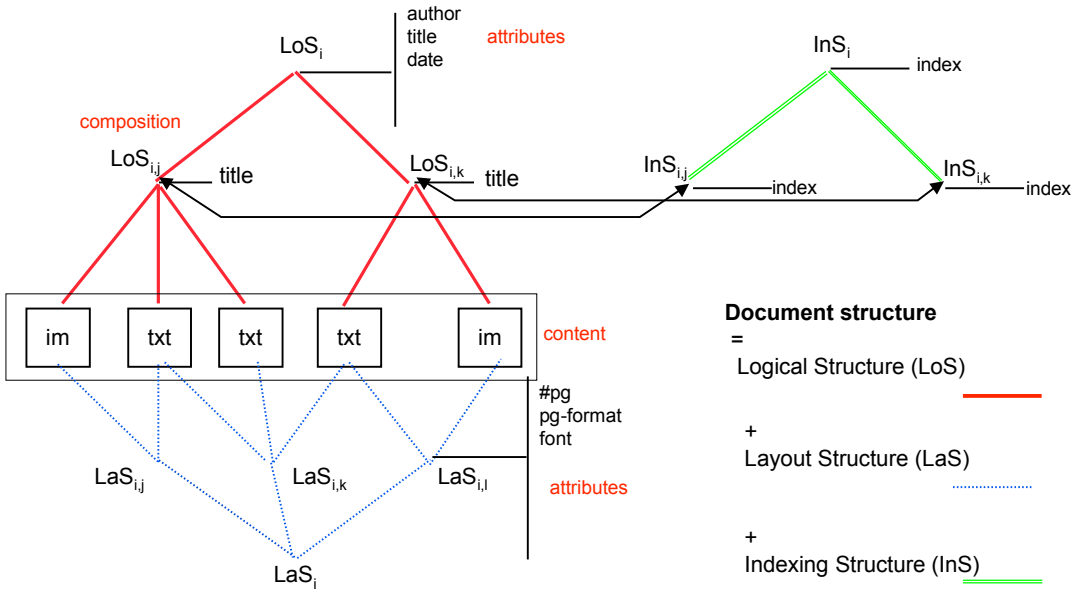
- **Notion of Structured Document**
- **The Impact of Structure**
- **Approaches**
  - Hypermedia
  - Passage Retrieval
  - Hierarchical Structures
  - Link structures
- **Conclusions**

Plan

- Notion of Structured Document
- The Impact of Structure
- Approaches
  - Hypermedia
  - Passage Retrieval
  - Hierarchical Structures
  - Link structures
- Conclusions

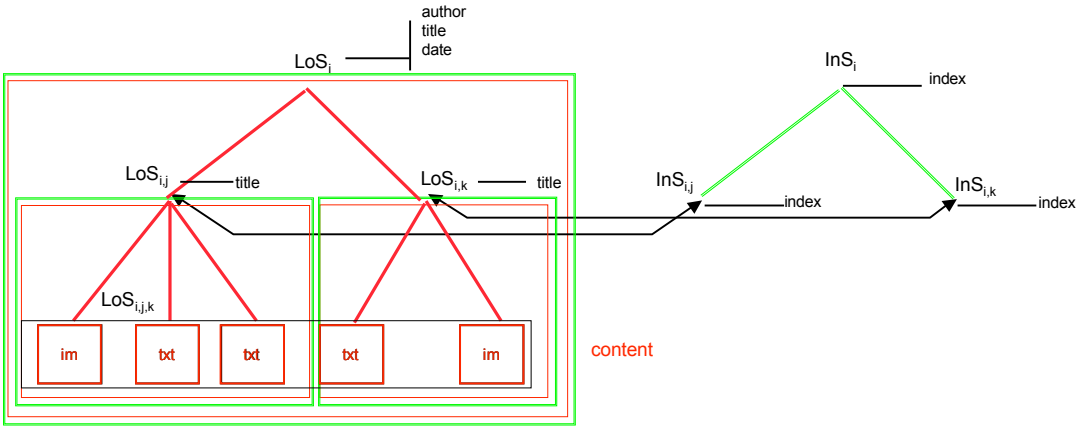
Notion of Structured Documents

- A complex notion, involving different facets: hierarchical structures



Notion of Structured Documents

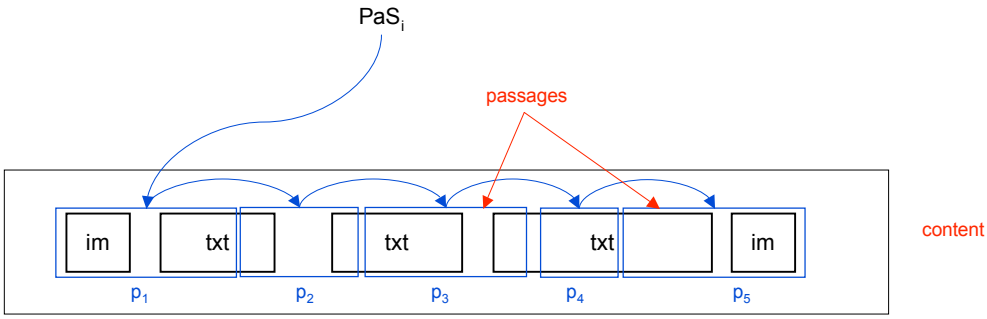
- **Hierarchical structures imply recursive definition of document content**



Trees and subtrees are indexed and retrieved independently

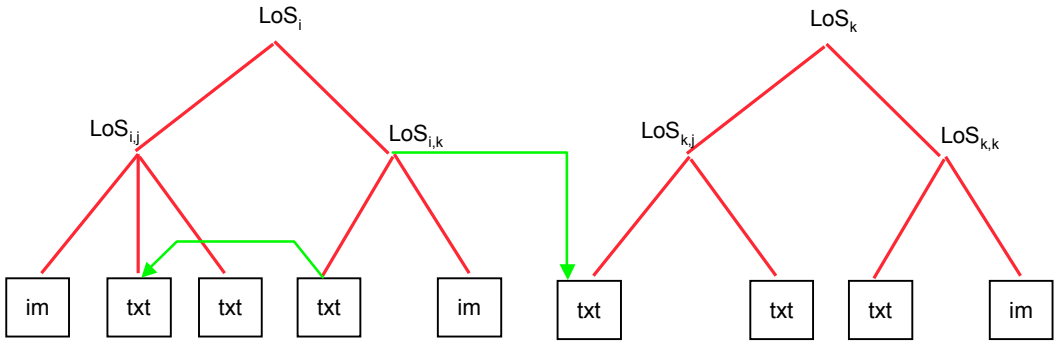
Notion of Structured Documents

- **A sequence of « passages »: linear structures**
  - Contain only textual data (may also include figure titles, legends..)
  - Various definitions of « passages » (generally ignores the logical structure)
  - **Passages are indexed and may be retrieved individually**



Notion of Structured Documents

- **Links and Hyperdocuments**



Document structure = logical structure + Layout structure + Indexing Structure

+

Browsing Links (internal, external) => HYPERMEDIA DOCUMENTS

Plan

- **Notion of Structured Document**
- **The Impact of Structure**
- **Approaches**
  - Hypermedia
  - Passage Retrieval
  - Hierarchical Structures
  - Link structures
- **Conclusions**

The Impact of Structure - Information needs

- **Information needs often involve structure**
- Variety of users' information requirements

On top of classical content-based requirements, information needs often include requirements about :

» **Attributes**
ex. *Novels written* by Hemingway

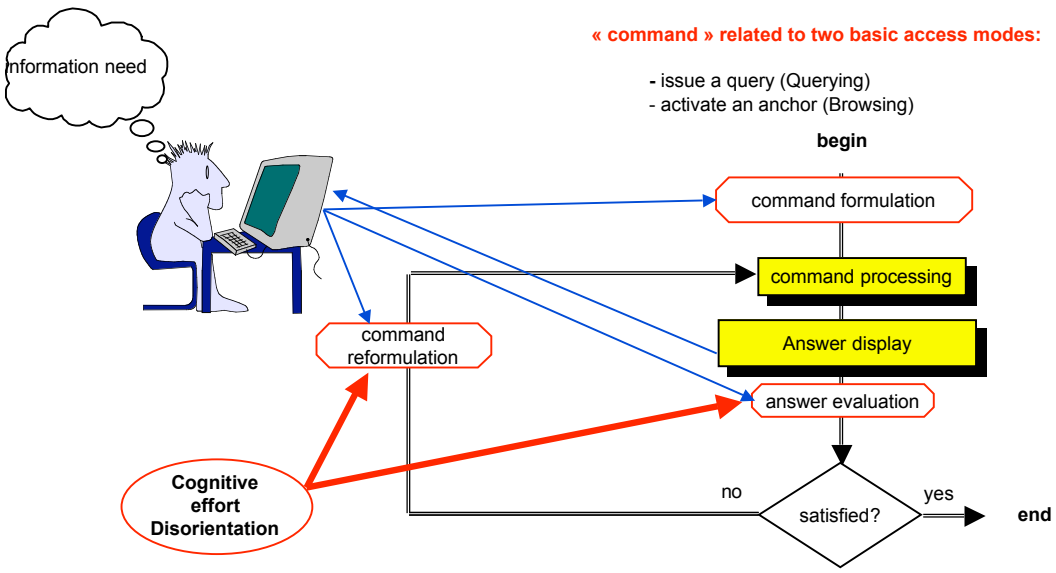
involved structural information : attributes type of doc, author

» **links**
ex. *A biography of Hemingway illustrated by* portraits *of the novelist.*

involved structural information :
- type of doc (text, image)
- component link between document and image

The Impact of Structure - Interactive IR

- **Interactive Retrieval & User tasks**



The Impact of Structure - Interactive IR

❖ Interactive Retrieval & User tasks

- Querying and Browsing are based on *explicit manipulation of structure* (although at different levels and for different purposes) :
  - querying : may use attribute values, logical structure, passages..
  - browsing : extensive use of links, structured knowledge
- Answer evaluation directly dependent on retrieved document length
- Query reformulation eased by topical specificity of (more focused) expected pieces of documents
- have complementary advantages & limitations

BROWSING

Cognitive effort:

- stacking paths and relevance judgements from previous steps

Disorientation:

- length of paths
- loops
- redundancy
- misleading paths

QUERYING

Cognitive effort :

- evaluating responses (Ri)
- properly reformulating query (Q i+1) from Ri

Disorientation :

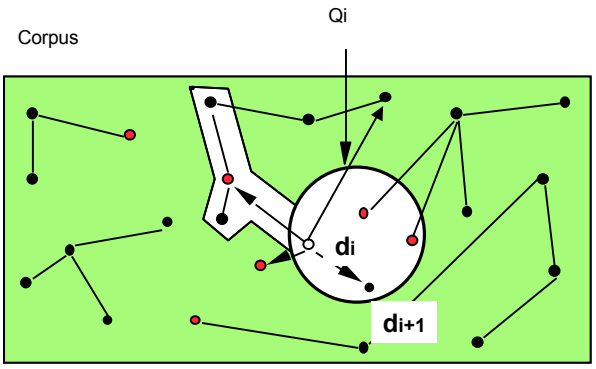
- length of responses
- bad ranking

➨ From the single point of view of ***interaction***, there is a need to make a proper use of structure

The Impact of Structure - Interactive IR

❖ Combining Querying and Browsing [YC97]

- Currently available on all search engines (Web)



Cognitive effort

- browsing = incremental,
- try and error, easier process
- may help query reformulation

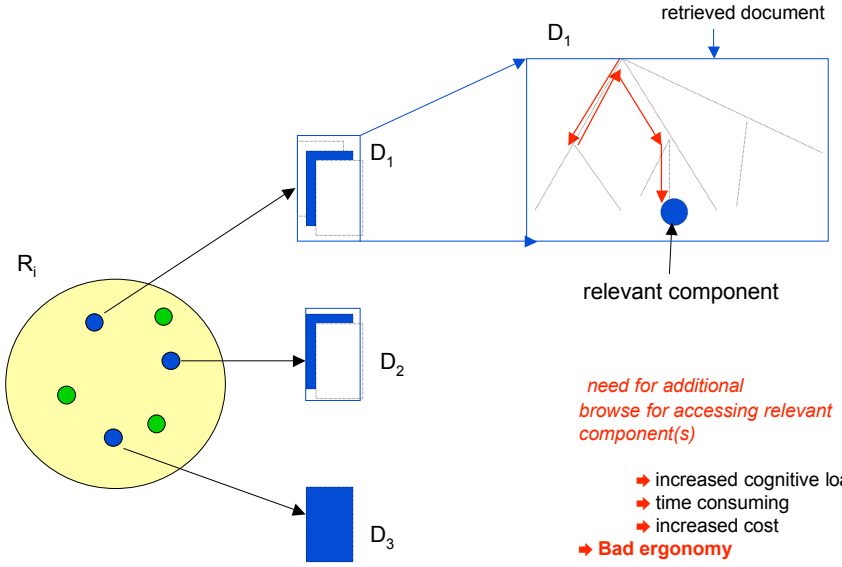
Disorientation

- querying effective for topic relocation In the document space

But still a need for models *integrating* querying and browsing !

The Impact of Structure - Interactive IR

- **Ergonomy - focusing on relevant components in *atomic* documents**



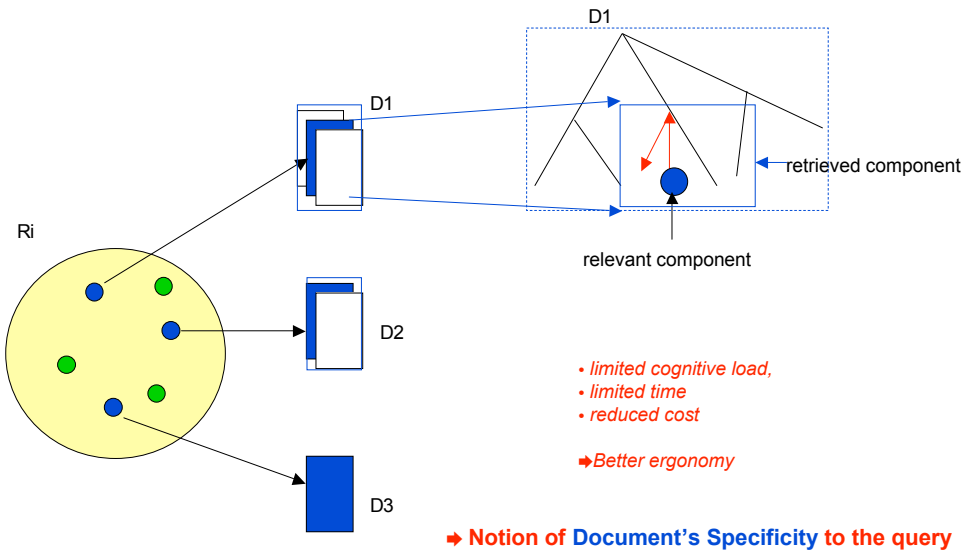
need for additional browse for accessing relevant component(s)

- increased cognitive load,
- time consuming
- increased cost

Bad ergonomy

The Impact of Structure - Interactive IR

- **Ergonomy - focusing on relevant components in *structured* documents**



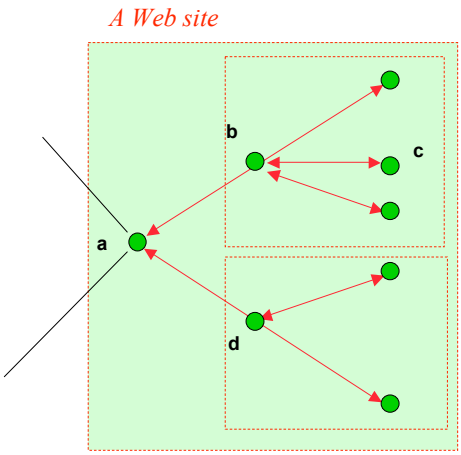
- *limited cognitive load,*
- *limited time*
- *reduced cost*

Better ergonomy

Notion of Document's Specificity to the query

The Impact of Structure - Interactive IR

❖ **Ergonomy - Disorientation Problems : the example of Web pages**



A web site has always an underlying hierarchical structure (its logical structure)

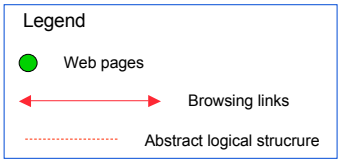
- **a** is the homepage of the Web site;
- **a** is also the root of the logical structure (a tree)

Links implement a **navigation** structure, not the **logical** structure of the Web site

➡The logical structure is not accessible

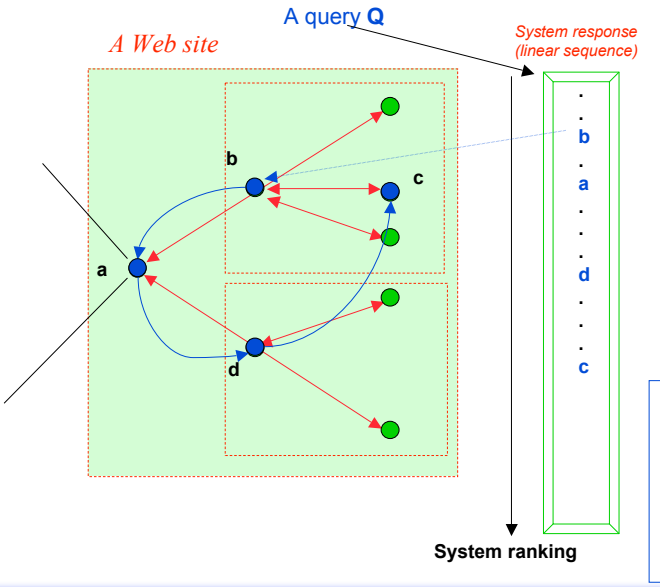
➡ page nodes are indexed separately

➡ **index of page a does not represent the content of subtree a (only of page a)**



The Impact of Structure - Interactive IR

❖ **Ergonomy - Disorientation Problems : the example of Web pages**

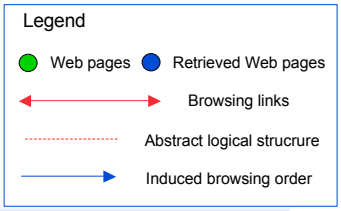


The logical structure is not visible, not accessible

➡ Browsing the retrieved pages in the proposed order (decreasing relevance of web pages) is most often not coherent with any consistent access order of the logical structure (prefix, infix etc.)

➡ redundant accesses (cognitive load, disorientation)

➡ the user has to re-engineer the structure to some extent

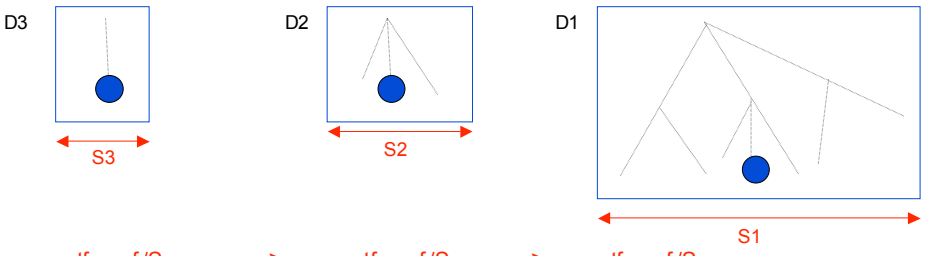


The Impact of Structure - Ranking

- **Incidence of structure on term weighting (tf-idf)**
  - Term weights affect the ranking of retrieved documents, hence Recall and Precision performances
  - Term weights depend on document size (**tf**-idf):

Normalized term frequency: $tf_{ij} = f_{ij}/S_i$ with S_i the size of the document

- Consequence with atomic documents containing similar relevant components



$tf_{3i} = f_i/S_3 \quad \geq \quad tf_{2i} = f_i/S_2 \quad \geq \quad tf_{1i} = f_i/S_1$

➡ **Any query retrieving D1 D2 D3 will end with $Rank(D_3) \leq Rank(D_2) \leq Rank(D_1)$**

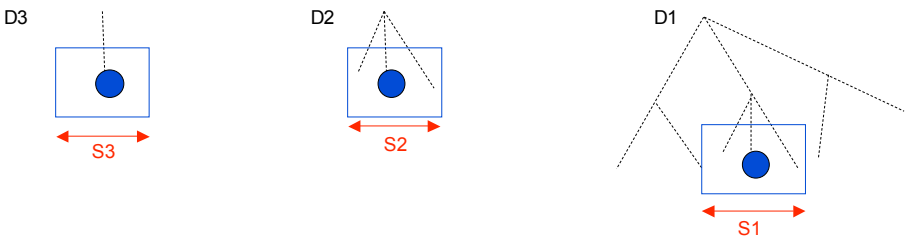
Document-size normalization may be not enough effective!

➡ **larger documents tend to be low-ranked**

➡ **bad Incidence on recall / precision**

The Impact of Structure - Ranking

- **Incidence of structure on term weighting (tf-idf)**



$tf_{3i} = f_i/S_3 \quad \approx \quad tf_{2i} = f_i/S_2 \quad \approx \quad tf_{1i} = f_i/S_1$

➡ **Any query retrieving D1 D2 D3 will end with $Rank(D_3) \approx Rank(D_2) \approx Rank(D_1)$**

Similar components (in size and content) tend to be closely ranked (whatever the size of embedding document)

➡ **Good Incidence on recall / precision**

The Impact of Structure - Conclusions

- **Due to its strong relationship with the notion of document content, document Structure has a major incidence on IR effectivenesss :**
  - As a basis for implementing *information access techniques* (querying, browsing)
  - As an information conveying its own semantics about *document relevance* (logical structure as a discourse structure, references, attribute values, impact of linked documents etc.)
  - As an information type allowing improvements in the *ergonomy of interactive retrieval*
- **Conclusion**
  - impact of Structure on *Interaction Performances* (cognitive load, disorientation, efficiency)
  - impact of Structure on *Retrieval Performances* (focusing)

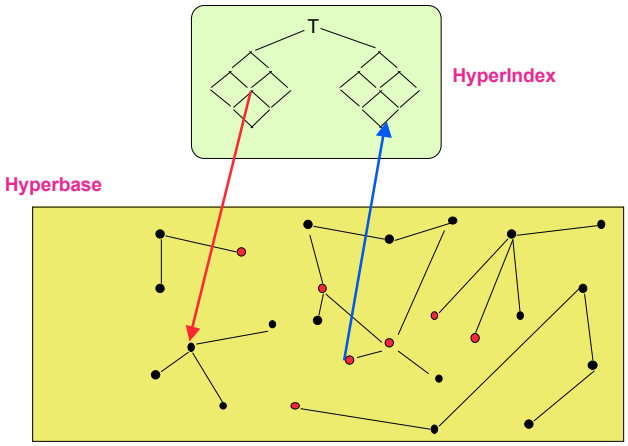
➡ **Need for better integration of Document Structure in Retrieval Models & Systems**
➡ **Need for Integration (more than combination) of Browsing and Querying capabilities**

Plan

- **Notion of Structured Document**
- **The Impact of Structure**
- ➡ **Approaches**
  - Hypermedia
  - Passage Retrieval
  - Hierarchical Structures
  - Link structures
- **Conclusions**

Approaches - Hypermedia

- **General structure** [ACG91] [VB90]
  - Two components
    - **Hyperindex**: a browsable knowledge base (keyword list, thesaurus, lattice of concepts etc.)
    - **Hyperbase**: a browsable base of documents
  - Two access functions:
    - **Beam down**
      Down from a concept to its class in the hyperbase
    - **Beam up**
      Up from a document to its Representative concepts
  - Building the hyperindex, the hyperbase classes, defining the beam down and beam up functions are aspects of the indexing process of an hypermedia retrieval environment



Approaches - Hypermedia

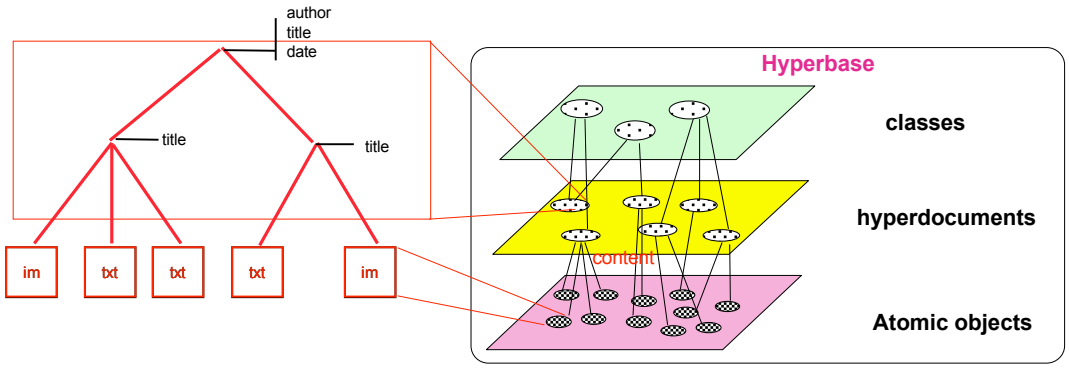
- **Integrating Browsing and Querying** [AK93], [AK95] [YC96a]
  - Extended IR model (hypermedia features) :
    - considering the document structure (including links, attributes) [CT89]
    - Extension to multimedia documents [LPZ93] [YC96b]
  - Extended Hypermedia Model : [FRE95]
    - Management of document content as explicite knowledge [NAN91] [LPZ93]
    - typed links [AK95]
    - weighted links
    - link construction

Need of a unifying model for
content
and
structural
knowledge
(eg. conceptual graphs)
[JFS84] [AK95]

Approaches - Hypermedia

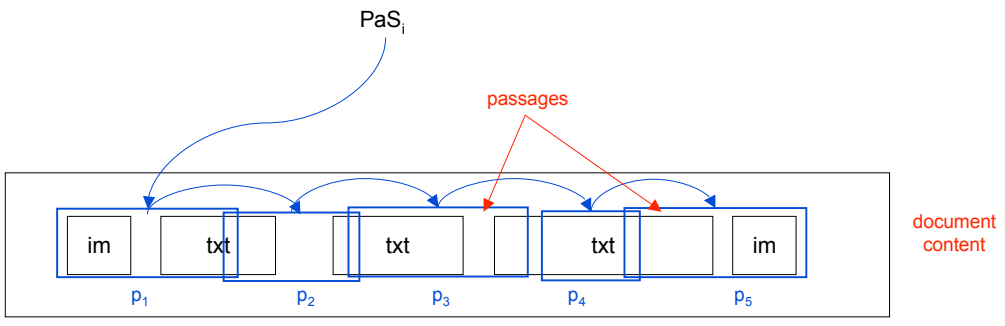
- **Integrating Browsing and Querying** [AK93]
  - Structuring the hyperindex (eg. Lattice of concepts)
  - Structuring the hyperbase [PG88]

Example: abstraction levels in the hyperbase



Approaches - Passage Retrieval

- **Document Structure**
  - Documents (mostly textual) are viewed as sequences of « passages »
  - The sequence may or may not cover the entire actual document content (text only or multimedia)
- **Great variety of passage definitions:** [SAL93]
  - Discourse: elements of the logical structure (sections, paragraphs etc.) [WIL94], [ZO95], [KAS97] [CAL94]
  - Semantic: sequences of words (overlapping or not), sentences etc. bounded by topic changes[CAL94] [MEL98] [MIT94] - Use of statistical / probabilistic approaches
  - Window: fixed length sequence of words (overlapping or not) [KIS01]

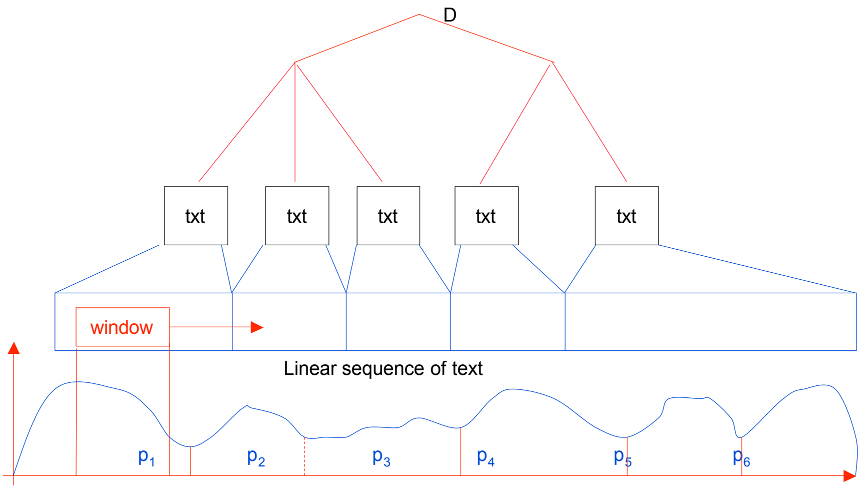


Approaches - Passage Retrieval

❖ **Passage Retrieval - segmenting texts into passages**

Example: fixed-length window [KIS01]

- document preprocessing: keep text only - eliminate stop words -apply word stemming
- Computing local term density within a fixed-length sliding window
- Discontinuities supposed to denote topicality changes are used as passage boundaries

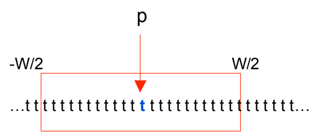


Approaches - Passage Retrieval

❖ **Example Evaluation of a term density in a document**

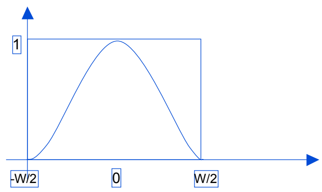
$t(p), 1 \le p \le L$ a term at position p in the document

t_i^q a *weighted* (w_i^q) query term



– Weight of a query term found in position *p* in the document

$$w(p) = \begin{cases} w_i^q . idf_i & \text{if } t(p) = t_i^q \\ 0 & \text{otherwise} \end{cases}$$

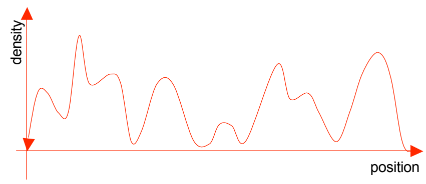


– Computation of term density, using a window of length W

$$dd(p) = \sum_{x=-W/2}^{W/2} f(x).w(p-x)$$

With f(x) a smoothing function on the window span:

$$f(x) = \begin{cases} \frac{1}{2}(1+\cos 2\pi \frac{x}{W}) & \text{if } abs(x) \le \frac{W}{2} \\ 0 & \text{otherwise} \end{cases}$$



Approaches - Passage Retrieval

- **Conclusion**

  - An active class of approaches

  - **advantages** :

    - simple to implement,
    - efficient,
    - can be effective (long documents)

  - **limitations** :

    - text only
    - incidence of document types (length)
    - decision making about segements boundaries not easy

  ➡ needs to be tuned to document types

  ➡ Strong impact of NLP techniques [HEA94]

  ➡ May be combined to other structural aspects (linksetc.) [CRI01]

  - **Question-answering** (an ideal for IR!) may be viewed as an extension[MOR99] of passage retrieval, based on deeper semantic analysis (involving NLP and AI). See [MOR99] See a vast amount of detailed bib And references to lestures on http://www.answerbus.com/bibliography/index.shtml

  An interesting - and successful - return to previous ideas (ex. [O'CON75] [O'CON80])!

Approaches - Hierarchical Structures

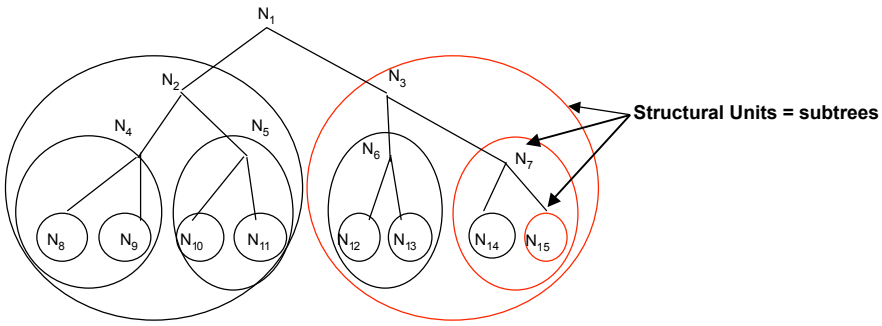
- An important case : hierarchical structure (textual documents, Web sites presentations, video..)

  – **Focus on logical structure and links**

  –Corpus = { structural entities } = **{trees}**

  –**Structural entities are related to document types (sections, subsections, paragraphs etc.)**

  ➡**Each unit indexed / retrieved independantly** (an old idea: [KER85])

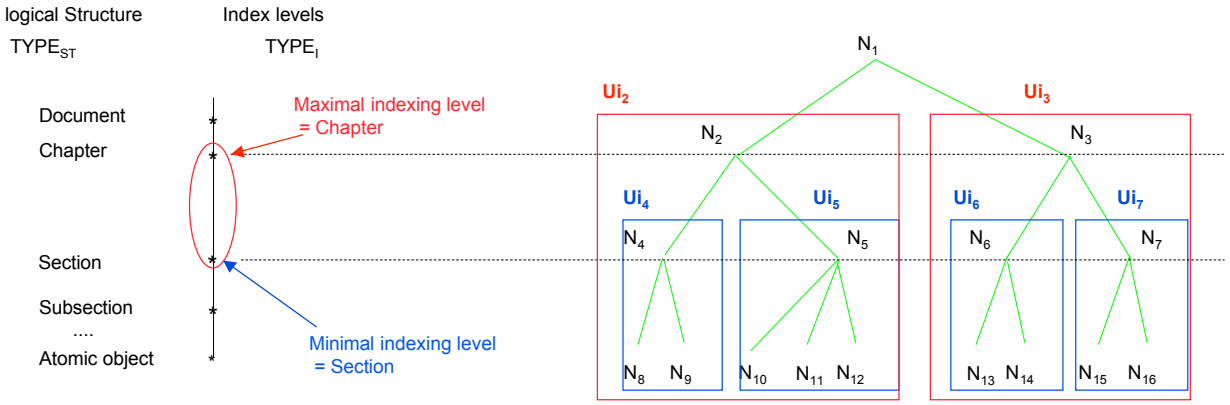


➡ **Relationship between structure - semantic content : individual indexing of structural units**

Approaches - Hierarchical Structures

❖ The Index Structure

- Structural Units (Logical Structure) vs Indexing Units (Index Structure)
- Maximal & Minimal indexing levels (informative units)
- Boundaries fixed at application level (document types, users' requirements)



Approaches - Hierarchical Structures

❖ The indexing process - Relationship between Structure and Content

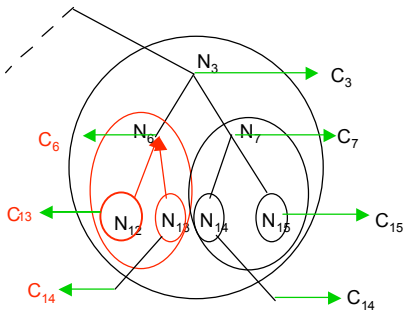
- Indexing structural components
  - **Goal** : to allow separate access to indexing units: need to ne able to Infer information related to document components (index, attribute values)
  - **Approach** : - Aggregation (⊕) as a ***bottom-up propagation*** of content from lowest-level, ***atomic***, components)

    - A unified vue based on ***attributes*** for storing index (*symbolic attribute*) and standard attributes (titles, authors, dates etc.)

$C_6 = \oplus(C_{13}, C_{14})$

$C_7 = \oplus(C_{14}, C_{15})$

$C_3 = \oplus(C_6, C_7)$



Three attribute classes :

- ◆ Dynamic Descending Attributes

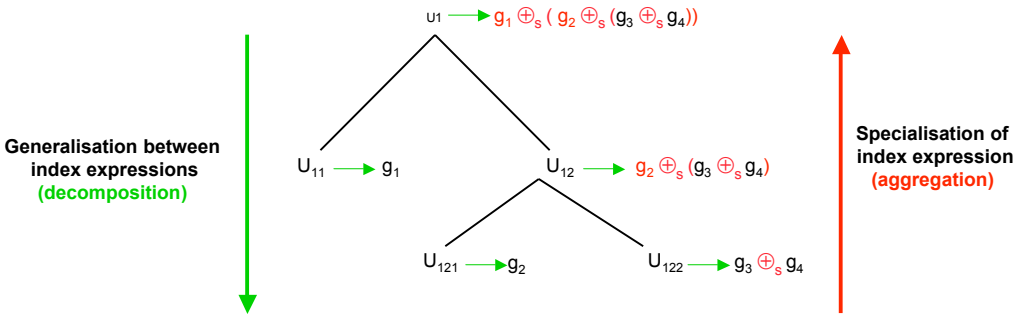
  Ex. Publication date
- ◆ Dynamic Ascending Attributes

  Ex. Symbolic attribute, author
- ◆ Static Attributes

  Ex. Title

Approaches - Hierarchical Structures

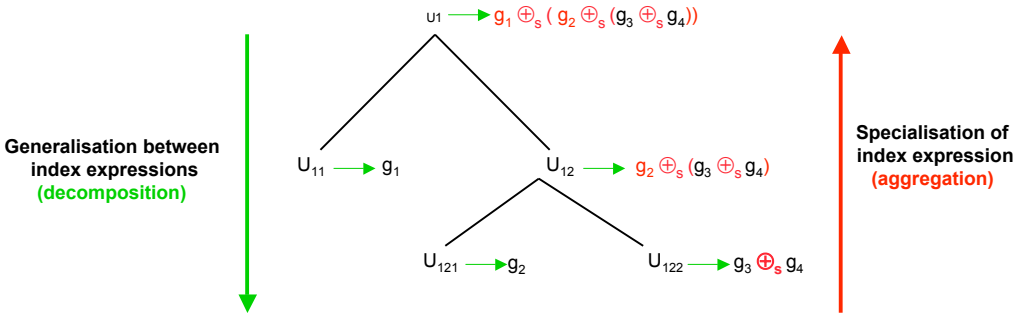
- **The indexing process - Relationship between Structure and Content**
  - Content Attribute : Dynamic, Ascending
  - Values of content attribute: expression of a symbolic language $L^{symbolic}$
  - Using operator $\oplus_{symbolic}$: ascending aggregation of values of content attribute for each index unit :



Generalisation of index expressions down the hierarchy :
an interesting property for
- controlling the composition within the hierarchy
- Implementing the notion of « focusing »

Approaches - Hierarchical Structures

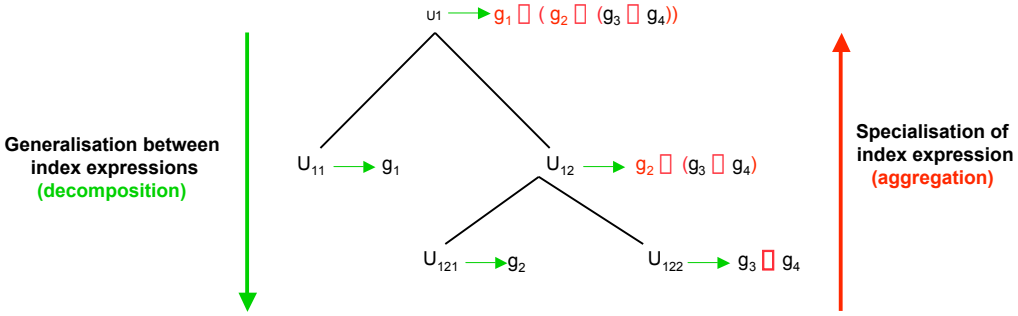
- **Properties of operator $\oplus_{symbolic}$**
  - Neutral element □ : g ⊕ □ = g
  - Reflexivity :g ⊕ g = g
  - Symmetry : g ⊕ f = f ⊕ g
  - Associativity : (g ⊕ f) ⊕ h = g ⊕ (f ⊕ h)



- **Properties defined for each attribute**
- **Determine a strategy for aggregating values**
- **fundamental for modeling relationship between content and structure**

Approaches - Hierarchical Structures

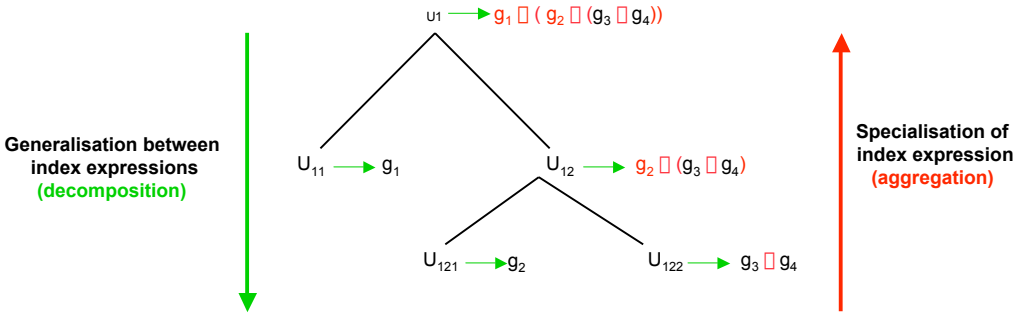
- **Example :** set indexing
  - $L^{symbolic}$ **= { index terms}**
    Ex: g1 = {'car', 'traffic', 'city'}
  - $\oplus_{symbolic}$ **is the union operator** $\cup$:



➡ **Could be extended to the aggregation of weighted terms**
$(\{a\}, w_a) \oplus_{symbolic} (\{b\}, w_b) = \{a \cup b\}, f(w_a, w_b)$

Approaches - Hierarchical Structures

- **Example :** boolean indexing
  - $L^{symbolic}$ **= boolean conjunctions of index terms**
    Ex: g1 = car ∧ traffic ∧ city
  - $\oplus_{symbolic}$ **is the** $\wedge$ **operator**



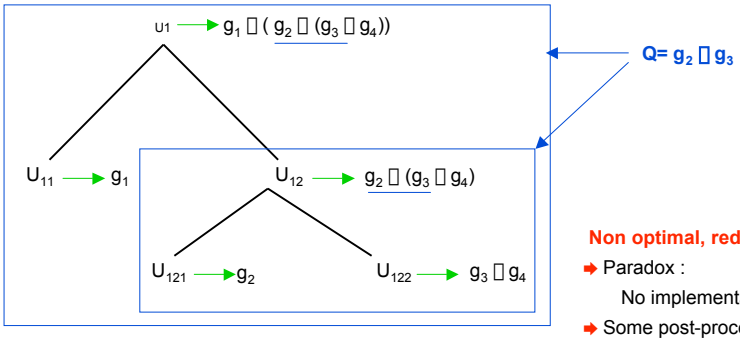
➡ **Could be extended to the aggregation of weighted terms**
$(a, w_a) \oplus_{symbolic} (b, w_b) = (a \wedge b, f(w_a, w_b))$

Approaches - Hierarchical Structures

- What strategy for retrieving indexed units ??
  - **A) Direct approach :**

    Example : boolean model

    direct application of D_i Q (D_i indexing units)



Non optimal, redundant answer

- Paradox :
  No implementation of the notion of « focus »!
- Some post-processing needed, depending on a particular definition of the notion of « focus »

Approaches - Hierarchical Structures

- What strategy for retrieving indexed units ??
  - **B) Soft approach :** logic-based models for IR [CJR86]

    Example : boolean indexing

    A two-step algorithm (Fetch & Browse)

    Based on the hierarchy of index expressions

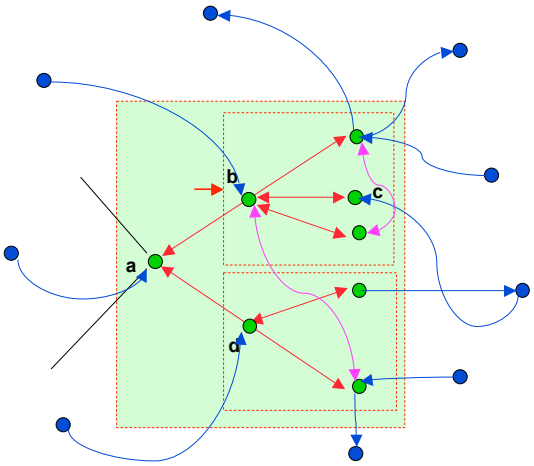
    1. Preselection of index units based on evaluation of
       D_i Q (standard retrieval based on **exhaustivity**) (C_1)
    2. Selection of ***minimal*** index unit
       Q D_i (an implementation of **specificity**) (C_2)

- If both conditions C_1 and C_2 are matched, D_i is an exact match for Q
- Else : the process recursively looks for the smallest units (ie. deepest in the hierarchy) such that:

Recursive case:	**Stop case**
D_i Q (exhaustivity requirement) and $\neg(Q$ $D_i)$ (specificity limitation)	$\neg(D_i$ $Q)$ (non exhaustivity) or $(Q$ $D_i)$ (minimal specificity)

Approaches - Link Structures

- **A general case**
  - Links may overlap with standard structures (logical)
  - Links may be (apparently) arbitrary, internal or external to the document



Questions

To which extent do links participate to the notions of

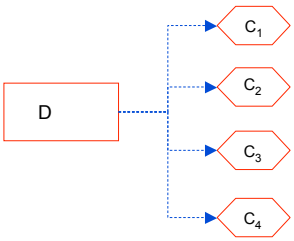
- document content?
- document relevance?

[CT89] [LDH92] [DR93] [AM01]

Approaches - Link Structures

- **Early related studies:**
  - **Impact of citations on indexing**

- Citations may be viewed as particular classes of links (even when purely symbolic and not browsable)



General approach

- use the information available in bibliographical references (titles) to improve the indexing of D
- extend to the process to actual referenced documents when possible

Results are quite controversial

- Irrelevant information introduced in the index of D
- There are various types of references denoting various intentions (various semantics)
- No definite(explicit) semantics associated to citation links in documents !

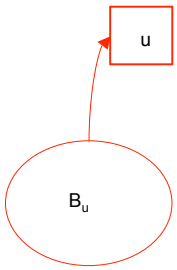
Approaches - Link Structures

❖ Links Contribution to Relevance

– The Google case [BMP99] [KL97]

- The notion of PageRank

- u a web page
- **Basic asumption**: the more page u has pages pointing to it, the more u is « important », or « authoritative »
- **The approach**:
  - Associate to u (on top of classical indexing), a **rank (named PageRank) : a number expressing the authority of the web page (part of the indexing process)**
  - Given a query Q, retrieving pages is done in three steps:
    - Use a conventional search engine with Q
    - Sort in descending order the retrieved pages according to their pre-computed PageRank
    - Display the retrieved pages in this order



➡ **Relevance enhanced by the notion of Authority (additonal to aboutness)**

➡ **A notion entirely related to structure!**

Approaches - Link Structures

❖ Links Contribution to Relevance

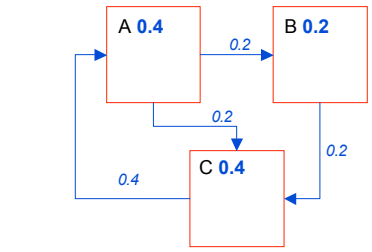
– The PageRank evaluation [BMP99]

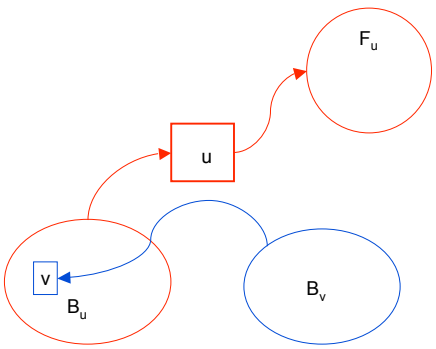
- u a web page
- F_u : the set of pages page u points to (outlinks) with $N_u = ||F_u||$
- B_u: the set of pages pointing to u (backlinks)

PageRank of page u : R(u) recursively computed from the ranks of pages pointing to it (backpages B_u) :

$$R(u) = c \sum_{v \in B_u} \frac{R(v)}{N_v}$$

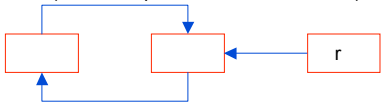
Example



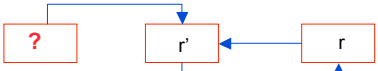


Approaches - Link Structures

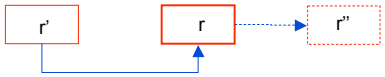
- **Links Contribution to Relevance** [BMP99]
  - **PageRank enhanced**
    - resolving Rank sinks (infinite loops of accumulative ranks)



    - resolving unrankable pages (no backlink)



    - resolving non contributing pages (no outlink - dangling references)



R" computed from r in a further iteration

Inclusion of Sources of Ranks E(u) in the evaluation

$$R'(u) = c \sum_{v \in B_u} \frac{R'(v)}{N_v} + cE(u)$$

Plan

- **Notion of Structured Document**
- **The Impact of Structure**
- **Approaches**
  - Hypermedia
  - Passage Retrieval
  - Hierarchical Structures
  - Link structures
- **Conclusions**

Conclusions & Perspectives

- Considering (ie. making explicit) structure leads to better retrieval performances in terms of interactive characteristics and classical retrieval performances (precision, recall)
- Considering structure may allow for a much powerful integration of browsing vs querying as complementary ways for retrieving information
- Considering structure may allow for the integration of various media in an integrated indexing/retrieval strategy
- Considering structure may help in improving focus / precision, a much needed improvement for searching the Web and any large data repository
- Considering structure implies better understanding of the core notions of document, user needs, document relevance
- Need to develop advanced models & derived IR systems
- **Fundamental problems for IR:**
  - **the relationship between content and structure**
  - **the relationship between relevance and structure**
  - ➡ **the semantics of links**

➡ Logic-based approaches still seems the most promizing (at least for theoretical models)

➡ Need to derive efficient technology from theoretical models

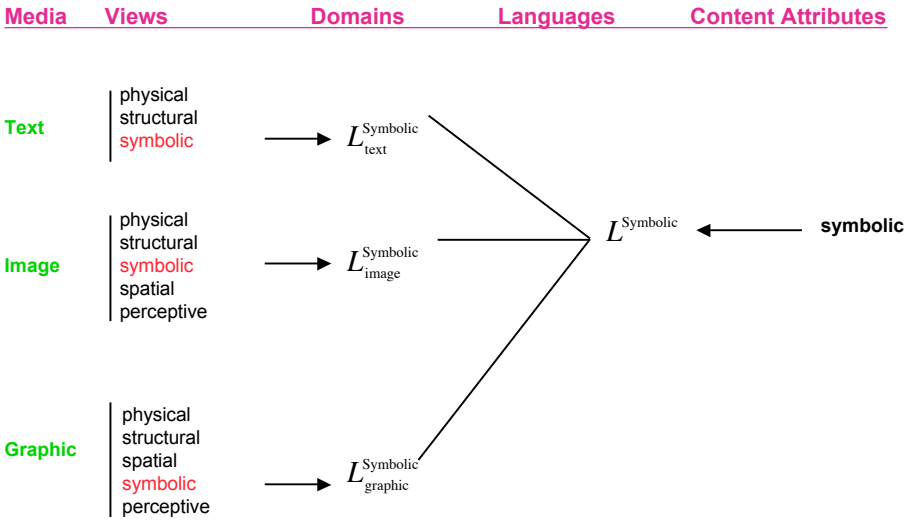
Approaches - Multimedia Documents

Indexing Multimedia Data based on attribute values

- Multivalued attributes
- domain values are expressions of a given language $L^{<\text{facet}>}_{<\text{media}>}$
- Used for integrating single-media models in the framework of structured, multimedia documents:
  - » Types of Content Attributes (facets) :
    - Physical
    - Symbolic
    - Structural
    - Spatial
    - Perceptive
- Associated to **views** of single-media data

Approaches - Multimedia Documents

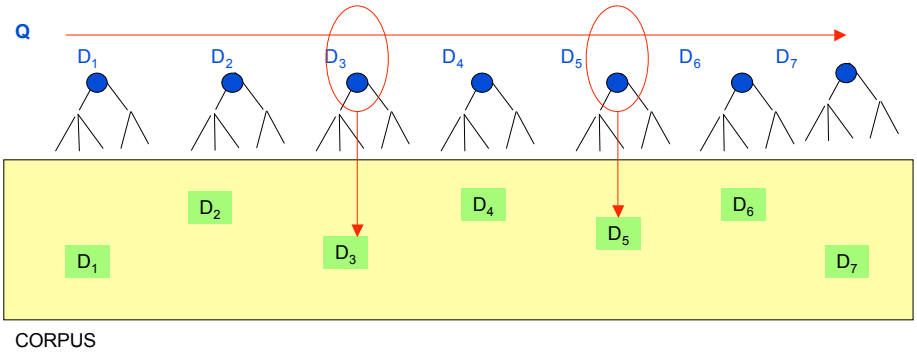
- **Combining Content Attributes for different media**



Approaches - Focusing in Hierarchical Structures

- **Example** : Fetch &Browse algorithm
  - **Fetch**

"horizontal" **preselection** of **documents** D_i satisfying D_i Q:



Approaches - Focusing in Hierarchical Structures

– **Browse**:

"vertical" selection of most specific units within the preselected documents (fetch) :

- Recursive Case : D_i Q and $\neg$(Q D_i)
- Stop Case : (D_i Q and Q D_i) $\Rightarrow$ result = D_i

or $\neg$(D_i Q) $\Rightarrow$ result = Father(D_i)