CENTRE NATIONAL
DE LA RECHERCHE
SCIENTIFIQUE

Institute for
Infocomm Research
A*STAR

Using Ontology Dimension and Negative Expansion In CLEF05

Jean-Pierre Chevallet (IPAL CNRS)
Joo Hwee Lim (I2R A-STAR)
Saïd Radhouani (CLIPS CUI)

Outline

- Characteristics of the collection
- The presence of query dimension
- How to take into account dimensions
  - Ontology focus
    - with negative weight
  - Ontology dimension importance
    - Dimension pre-filtering
- Results

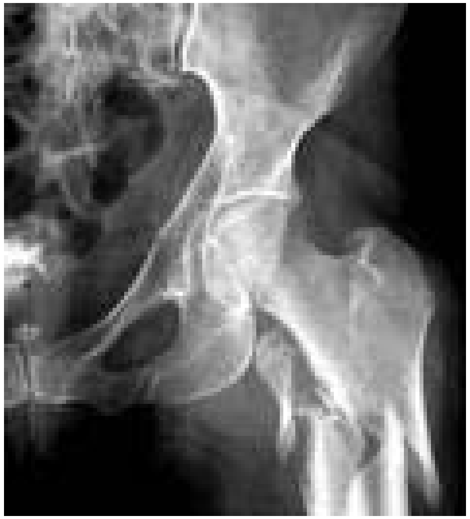
ImageCLEF 2005: Medical Retrieval Task

- 50,026 medical images from 4 collections:
  - Casimage: Radiology and pathology
  - Mallinckrodt Institute of Radiology (MIR): Nuclear medicine
  - Pathology Education Instructional Resource (PEIR): Pathology and radiology
  - PathoPIC: Pathology
- Annotations in XML format. The majority in English but a significant number also in French and German, with a few cases that do not contain any annotation at all
- Topics/Queries are expressed in 3 languages (English, French, German) + example images (all positive examples except one query with negative example)

Query dimensions

- Show me x-ray images with fractures of the femur.
- Zeige mir Röntgenbilder mit Brüchen des Oberschenkelknochens.
- Montre-moi des fractures du fémur.

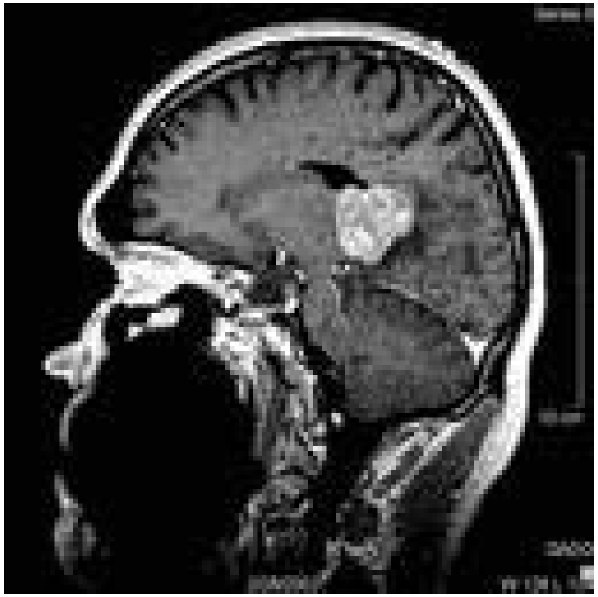
Modality **Pathology** **Anatomy**



Query with modality refinement

- Show me sagittal views of head MRI images.
- Zeige mir sagittale Ansichten von MRs des Kopfes.
- Montre-moi des vues sagittales d’IRMs de la tête.

Modality ***Pathology*** **Anatomy**

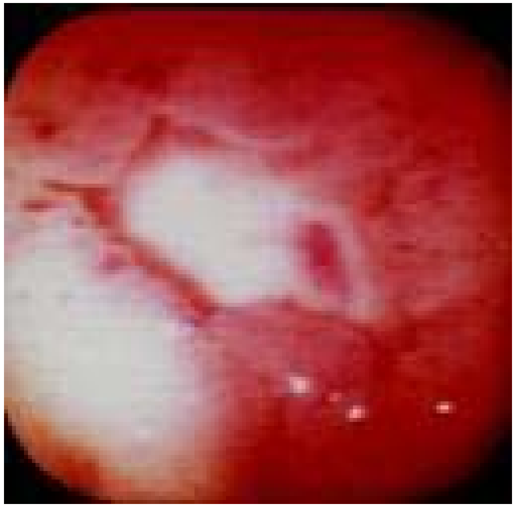


- No explicit Pathology

Queries more “semantic”

- Show me images showing peptic ulcers or part of it.
- Zeige mir Bilder eines Magengeschwurs.
- Montre-moi des images d’ulcères de l’estomac.

Modality **Pathology** ***Anatomy***

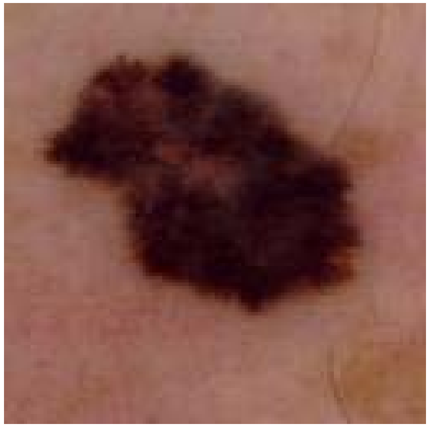


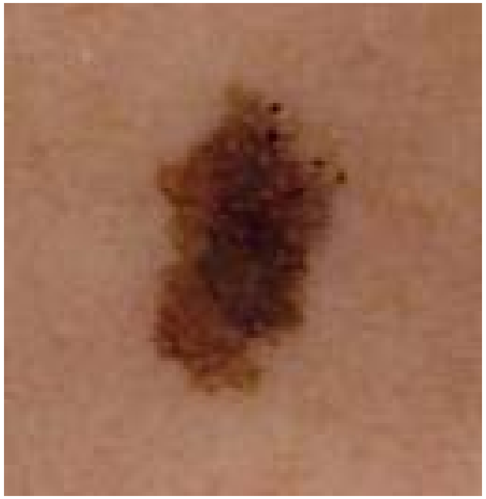
Anatomy is **explicit** in French, and German

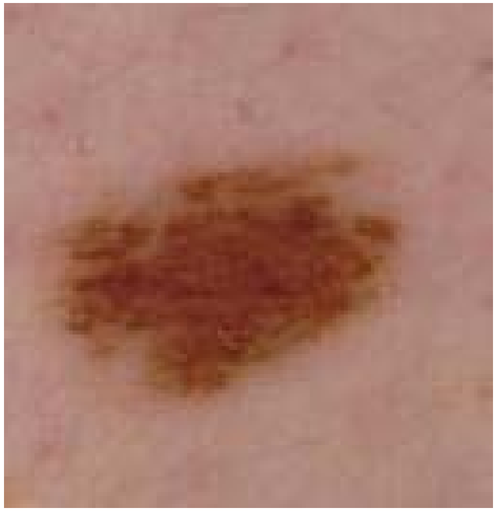
Queries with negative feedback

- Show me any photograph showing malignant melanoma.
- Zeige mir Bilder bösartiger Melanome.
- Montre-moi des images de mélanomes malignes.

Modality **Pathology** ***Anatomy***







Implicit Anatomy (skin)

□□ In the case of **Precision Oriented Retrieval**

Very Special Queries

- Show me a guitar with a cancer
- Show me an old X-ray tool from Middle-Age
- Show me the face of a very nice guy

(All made at Vienna)

Precision oriented Retrieval

- What is a precision oriented Retrieval ?
  - Documents corpus on a **restricted** domain
  - User are **specialist** of this domain
  - **Precise** need (strong focus)
  - **Short** list of good quality results
  - Precision is preferred to recall
- What does it implies ?
  - Document are consistent in themes
  - Use of **terms** : words that belong to a **terminology**
    - Terminology: set of technical terms form a domain
  - Queries have **dimensions** related to an ontology

Ontology & Dimension

- "An ontology is a formal explicit specification of a shared conceptualization" [Gruber 93]
  - Formal: machine readable
  - Explicit: definition of types and constraints
  - Shared: group of people, reuse
  - Conceptualization: abstract model of some phenomenon, selection of the related concepts.
- We call ***Ontology dimension*** the first level of a domain ontology
  - Ex: from MESH (the three levels we use)
    - Anatomy [A]
    - Diseases [C] (Pathology)
    - Analytical, Diagnostic and Therapeutic Techniques and Equipment [E] (Modality)

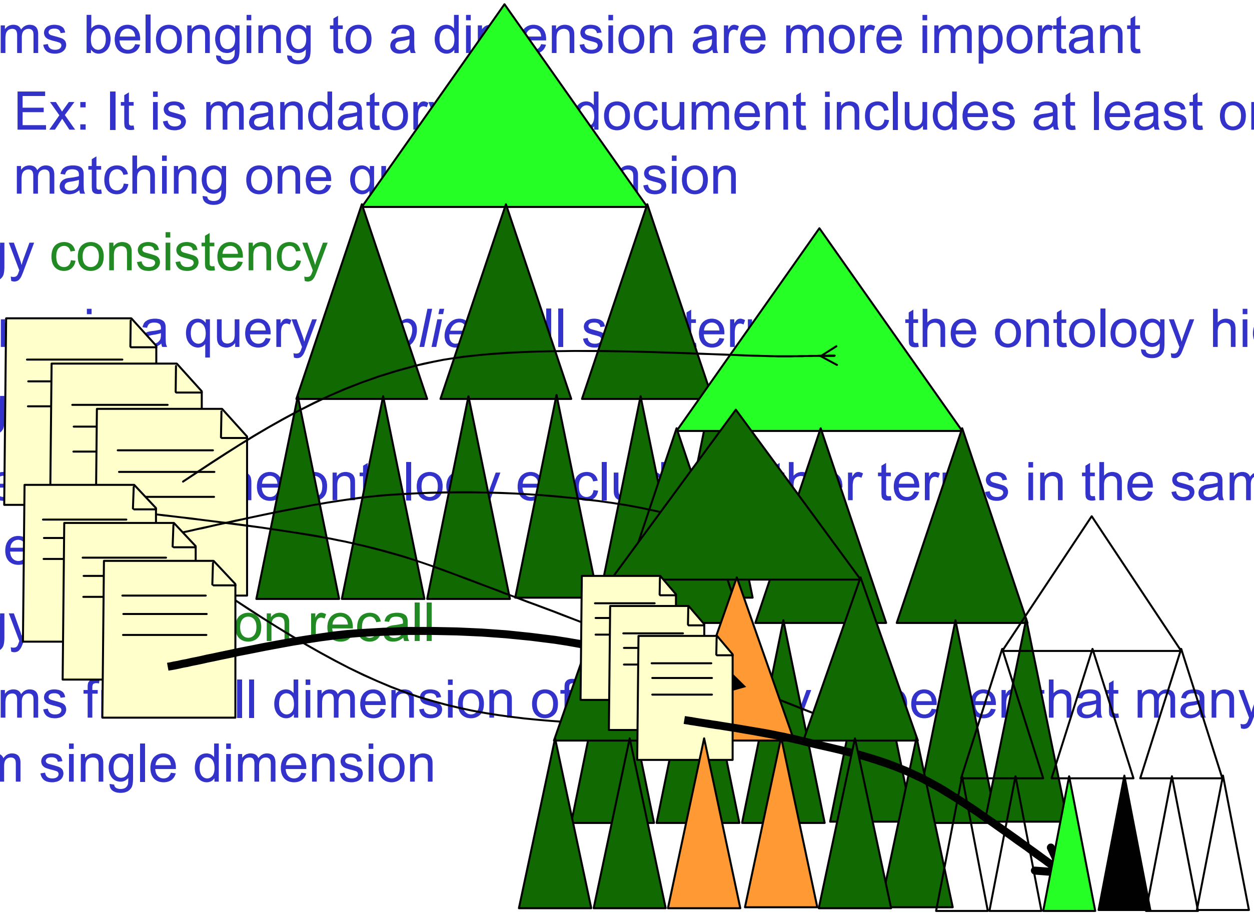
How to use dimensions ?

1. Dimension **definition**
   - Refers to an existing ontology (ex: MESH or UMLS)
2. Dimension **extraction**
   - Selection of the correct term
   - Related to Word Sense Disambiguation
3. Dimension **inclusion** in an IRS
   - Different “class” of terms
   - Need filtering among dimensions

- Somme possible solutions
  - Splitting queries on dimensions
  - Changing weight

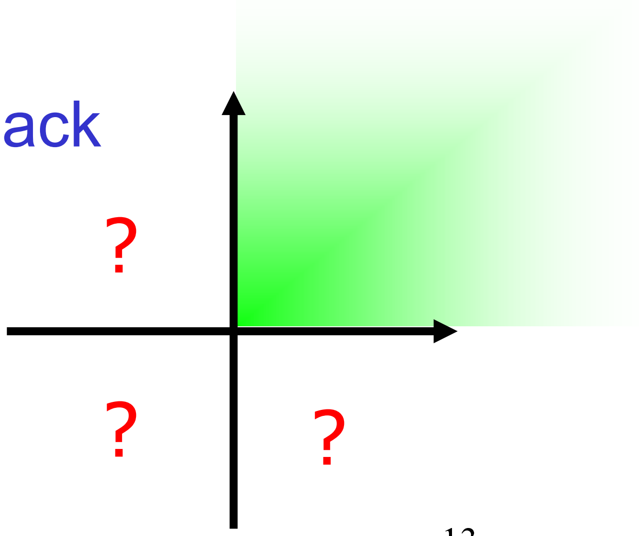
Some Hypothesis

1. Ontology dimension
   - First level on the hierarchy are meaningful dimensions
2. Ontology dimension importance
   - Terms belonging to a di[illegible]ension are more important
     - Ex: It is mandator[illegible] document includes at least one term matching one q[illegible]nsion
3. Ontology consistency
   - Ter[illegible] a query [illegible]*plie*[illegible]ll s[illegible]ter[illegible] the ontology hierarchy
4. Ontolog[illegible]
   - One [illegible]e ont[illegible]lo[illegible] e[illegible]clu[illegible]r ter[illegible]s in the same dime[illegible]
5. Ontology [illegible]on recall
   - Terms f[illegible]ll dimension of [illegible]e[illegible]hat many terms from single dimension



Ontology focus & VSM

- One term in the ontology excludes other terms in the same dimension
- Idea: using the normal Vector Space Model
  - Positive weight for terms that appears
  - Negative weight for terms on the same dimension that **does not appear**
- **Negative weighting** : seldom used
  - can appears during relevance feedback
- Is the VSM still consistent ?



VSM: inner product

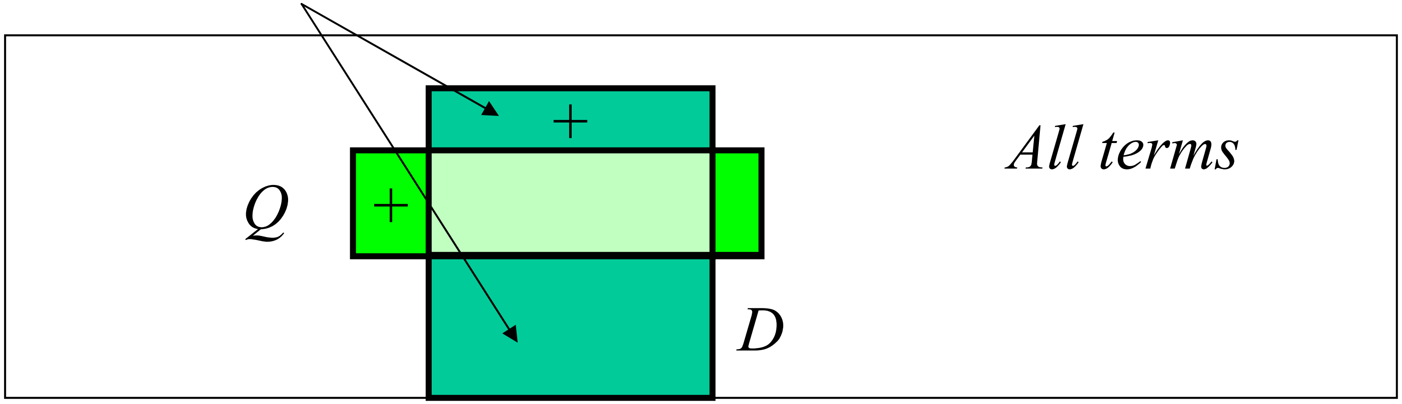
- Vector Space Model: matching based on inner product

  D □■■■□□□□□□□□□□■■■■■□□■■□■□□■□□□□□□□■■■□□

  Q □□□□□■□□□□■■□□□□■□□□□□■□□□□□□□□□□■□□□□□□

  A □□□□□□□□□□□□□□□□■□□□□□■□□□□□□□□□□□□□□□□□

  - For binary, it is the size of the **intersection** term set, with weighting still related to the intersection
  - Only query term participates to the matching
- Extra "**inconsistent**" document terms have no effect



VSM: lake of expressiveness

- Every terms at the same level
  - The classical “bag of word” problem
  - No way to force presence of terms
- No negation
- Only terms in the query participate on the matching
  - Classical solutions:
    - query expansion
    - Pseudo relevance feedback
- Expected behavior:

  “a change in the query implies a change in the matching”

Let’s have a look at the logical side …

Logical IR model

- Come from ideas of Keith Van Rijsbergen

  Relevance is expressed by *P(D→Q)*
- Ex: the “logical interpretation model”
  - Given a set of terms, an **interpretation** is a mapping to Boolean values
  - Formula are associated with **set** of interpretations
  - Logical **implication = inclusion** of interpretation
- In IR, document are expressed using only one interpretation
  - Meaning : *true* for terms relevant to document *D*
- Queries are expressed by a **set** of interpretation
  - Ex: $a \wedge b$ is associate to all interpretation where *a* and *b* are true

Logical Interpretation of VSM

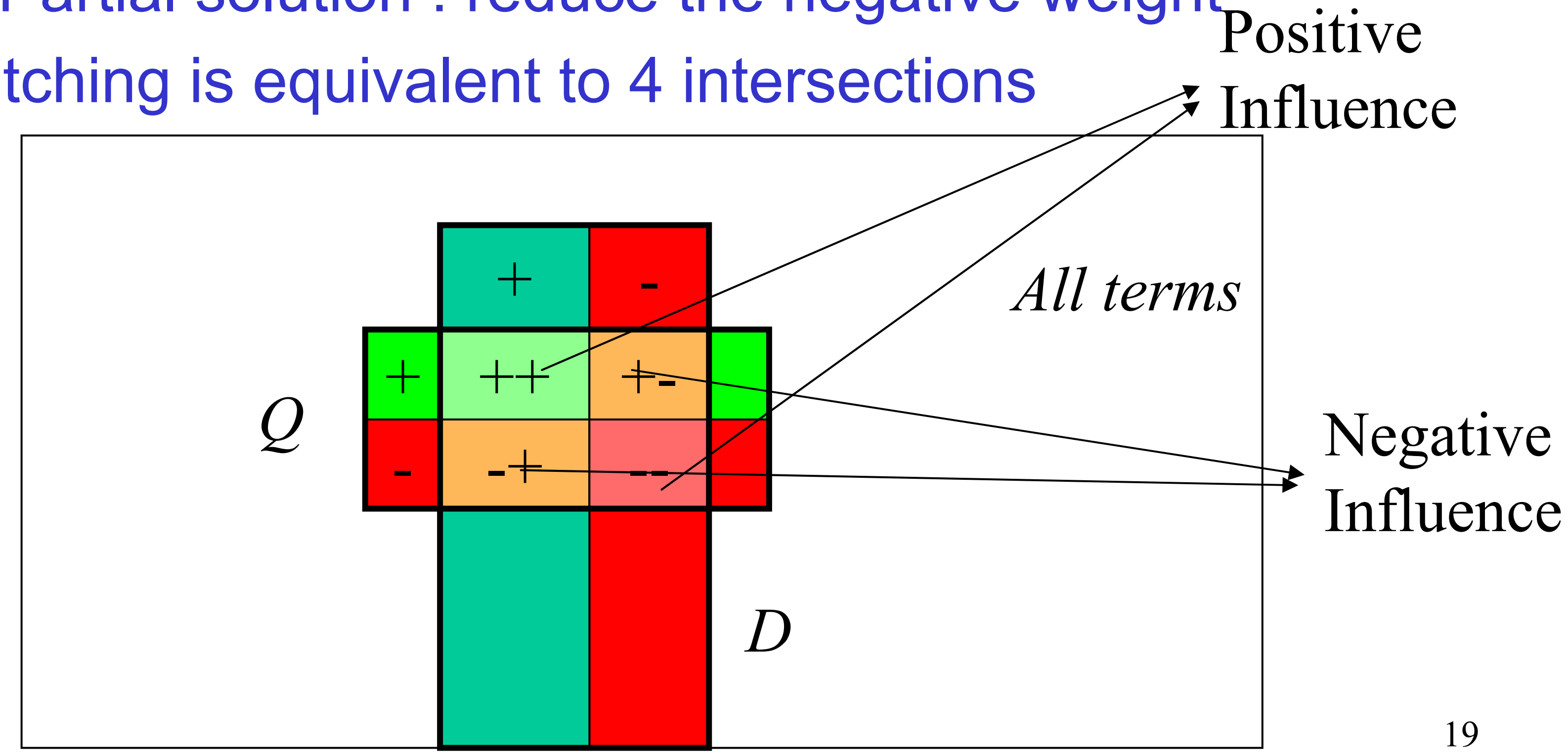
- Query for VSM are equivalent to:
  - **Disjunction** formula when correct matching for *non null inner product*
  - **Conjunction** formula when correct matching for inner product *equal to the term set query size*
  - **Something fuzzy** in-between (inclusion)
- Logical modeling has more power on the query side
  - Document: only one interpretation
    - Sort of closed word assumption
  - Query:
    - Several interpretation in boolean
    - Only one possible in VSM
  - Negation
- **New interpretation of VSM**

New interpretation schema

- New interpretation of terms
  - A term in relevant : positive value
  - A term is not relevant : negative value
  - No information on this term : null value
- Keep the use of inner product
- Query for VSM are then equivalent to:
  - **Conjunction** formula when correct matching for inner product *equal to the size of non null query terms set*
  - **Disjunction** formula : for a non null *positive influence (see next)*
- If no null term in index, then all query terms influence the matching

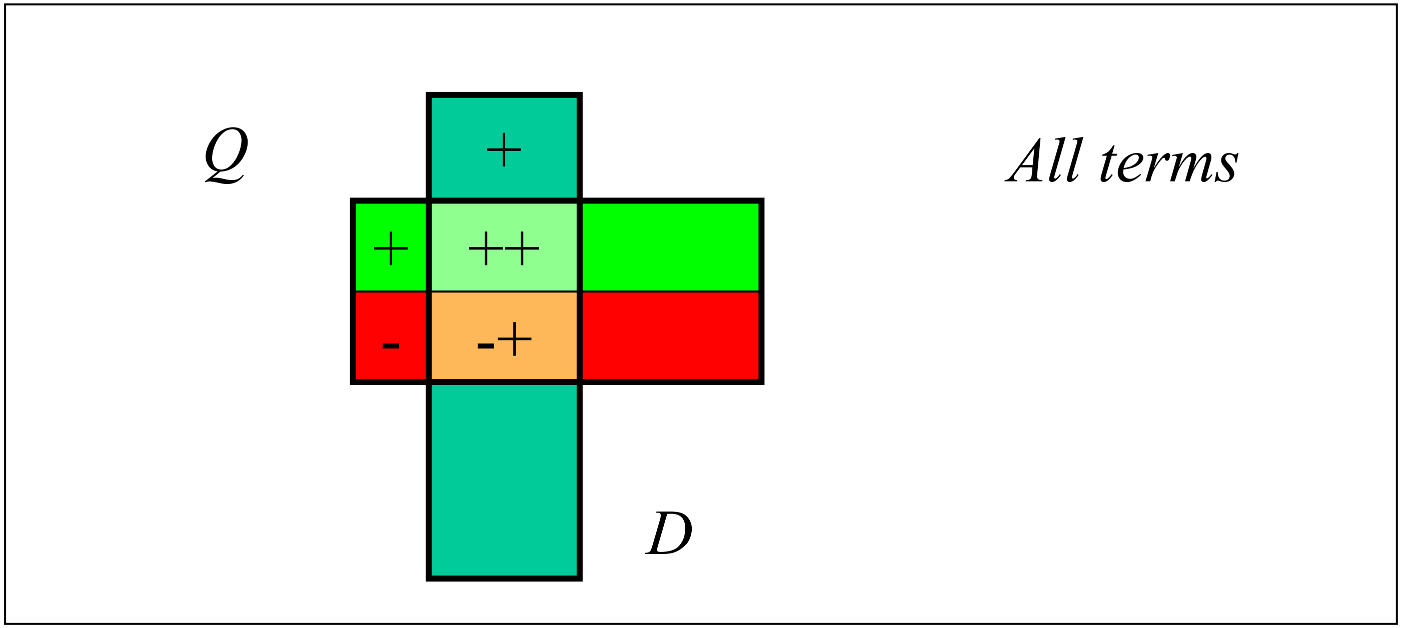
In practice

- Enlarge the indexing matrix
  - More computation to perform, inverted file less effective
- Negated terms may have a major role in matching results
  - Partial solution : reduce the negative weight
- Matching is equivalent to 4 intersections



Simplification

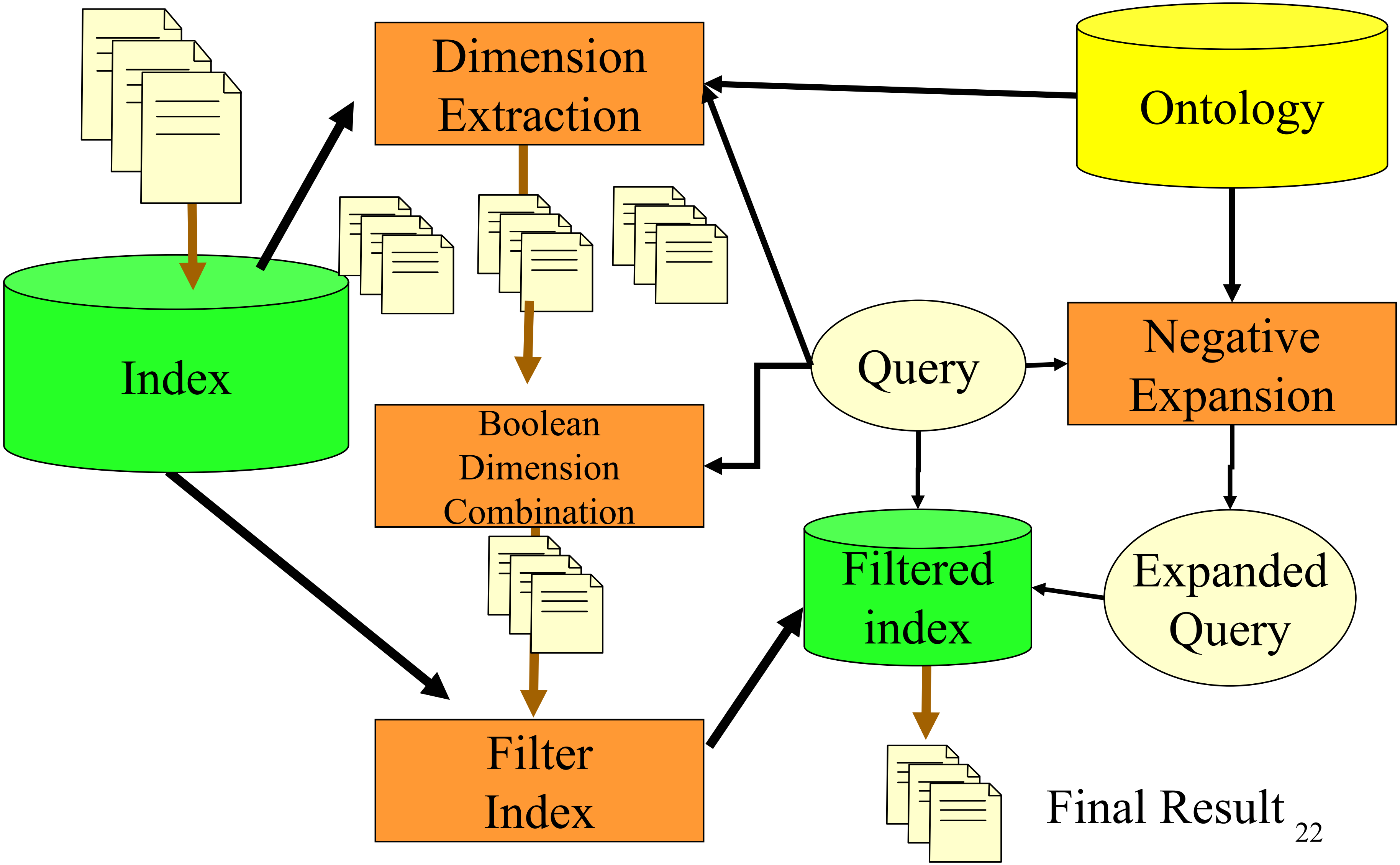
- We only have positive terms in documents
- We reduce the importance of negative terms
  - Negative expansion on the basis of the ontology
  - Equal weight distribution on added negated terms



Second approach
Dimension Filtering

- Ontology dimension importance
  - Terms belonging to some dimensions are more important
- Split the initial query
  - Each sub query is addressing one ontology dimension
  - A *mapping* query on ontology dimension (term set intersection)
- Use Boolean expression for dimension combination
  - Acts as a Boolean dimension filter
- Use classical VSM and weighting on this document filtered subset

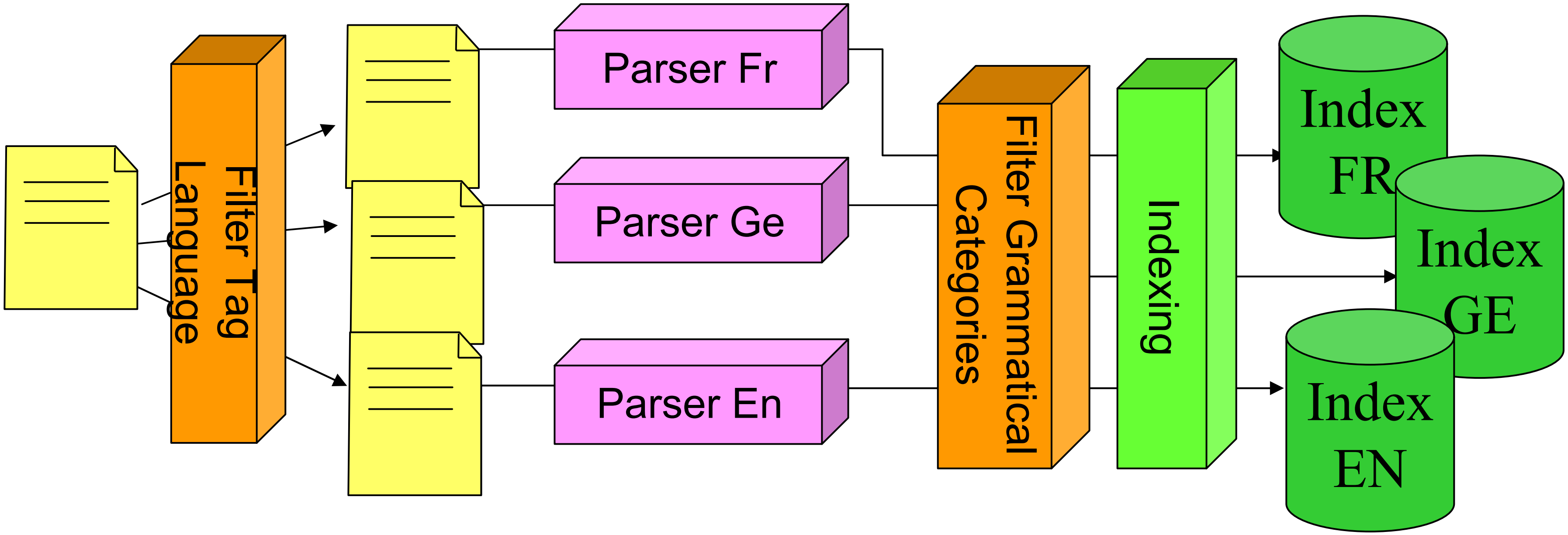
Global Framework



Indexing process

- Using XIOTA: XML based Information Retrieval Tool
- Correction of XML coding errors
- Automatic Reconstruction of document XML structure in MIR
  - Ex: “Brief history”
- Filtering fields to be indexed
  - Ex: PathoPic : Diagnosis Synonyms Description
- Part of Speech tagging
  - no stop words, “TreeTagger”
- Term selection: filtering on POS
  - Only nouns adjectives and abbreviations
- One single collection for all documents
- Same treatments to queries

Text extraction and indexing



- 3 sorts of semantic dimension : anatomy (hand, brain, etc), modality (MRI,Xray,etc),

Querying

- Using classical *ltc* weighting schema
- Querying each collection language
- Merging the result: take the maximum of RSV
- Selecting a very small subset of the ontology
  - First attempt to use dimensions
  - Reduce computation complexity
- Negative expansions of all queries using ontology
- Select one dimension per query
  - Filtering the index using dimension

Results

Negative Exp.	Dimension filtering	MAP
no	no	17,25%
yes	no	17,32%
no	At least one dimension*	19,64%
no	At least one particular dimension	20,75%
yes	At least one particular dimension	20,84%
yes	Anatomy	20,85%
yes	Anatomy & Pathology	21,39%

*Order:
Anatomy
Pathology
Modality

} Official runs

Image Indexing
A Structured Learning Approach

- 39 **visual keywords** (e.g. mri-head-brain, photo-skin, xray-lung-opaque) learned from 1460 cropped image patches
- Training images: 158 (0.3%) from the 4 test collection and 96 images from Web
- Support vector machines with RBF kernels
- Indexing: multi-scale detection of VK, reconciled and aggregated into local semantic histograms as compact indexes
- Support both similarity-based and semantic-based queries

(See poster of Lim Joo Hwee)

Fusion: Text + Image

- Fusion at retrieval level: based on query results for each query :
  - **Linear normalization** of text and image output (RSV: return status value)
  - Fusion schemes attempted:
    - Maximum of RSVs from 2 lists
    - Average of RSVs from 2 lists (intersection)
- Fusion at index level: enhance image index from text and vice versa (not tested)

Results with images

Fusion Method	Neg. Exp	MAP
Maximum	no	23,12%
Maximum	yes	23,25%
Intersection + Average	no	28.19%
Intersection + Average	yes	28.21%

Official runs

Conclusion

- CLEF05 multilingual image collection:
  - Very difficult task: more images, more languages, more “precise” and “semantics” queries
- Classic vector space model only fail :
  - documents too different in size (from 3 to 5000 words)
  - Precise semantic field in the query (anatomy, modality, etc)
- Importance of ontology dimension
- Dimension filtering more efficient than negative weighting
  - Hypothesis: ***Dimension importance*** better than ***ontology focus***
- Visual and text: a good complementary
  - Text: closer to the meaning, e.g. specific terms
- Some queries are easier by image contents, others are more appropriate using text.

Thank You !