

User Centered and Ontology Based Information Retrieval System for Life Sciences



Sylvie Ranwez

Mohameth-François Sy

Jacky Montmain

Michel Crampes

Vincent Ranwez

LGI2P Research Center / Ecole des Mines d'Alès, France ISEM - CNRS / Montpellier II University, France



Context and objectives

Ontology based information retrieval

Relevance calculus between a document index and a query

Similarity between two concepts

Relevance of a document with respect to a concept

Relevance of a document with respect to a query

Results visualization

Conclusion et perspectives

Context: *usual information retrieval engine*

The screenshot shows the Yippy search engine interface. At the top, there is a navigation bar with links for 'Search', 'Register', 'Download', 'Privacy: ON / OFF ?', 'Sign In', 'Contact Yippy', and 'More'. Below this is a search bar containing the text 'semantic web applications' and a 'Search' button. To the right of the search bar are links for 'advanced preferences'. Below the search bar, there is a sidebar on the left with a 'remix' button and a list of categories: 'All Results (139)', 'Tools (22)', 'RDF (14)', 'Project, Australia (7)', 'Building Semantic Web (9)', 'Semantic Web Applications and Perspectives (9)', 'UK (6)', 'Engineering (7)', 'Programming (7)', 'Platform (6)', and 'Apps, 10 Semantic (3)'. Below the categories is a 'find in clouds' search box and a 'Find' button. At the bottom of the sidebar is a 'Yippy Approved' logo with the text 'has a search engine'. The main content area displays 'Top 139 results of at least 2,390,000 retrieved for the query semantic web applications (details)'. Below this is a list of search results, each with a title, a brief description, and the source URL. The results include: 1. '10 Semantic Apps to Watch' from readwriteweb.com; 2. 'SWAN (Semantic Web Applications in Neuromedicine) Project' from swan.mindinformatics.org; 3. 'Semantic Web - Wikipedia, the free encyclopedia' from en.wikipedia.org; 4. 'Iowa State University - Center for Computational Intelligence, Learning, & Discovery' from www.cild.iastate.edu; 5. '10 Semantic Apps to Watch - One Year Later' from readwriteweb.com; 6. 'W3C Semantic Web Activity' from www.w3.org; 7. 'Semantic Web Applications and Tools for Life Sciences' from www.swat4ls.org.

Boolean search

+ Results are easy to understand

- Exact terms matching
- Number of results
- Rough measurement: "match" or "does not match"
- Limited interaction
- Aggregating operators are not used (AND, OR...)

⇒ Hard to grasp even with clustering

Context: *information retrieval based on a concepts hierarchy*

PubMed.gov
U.S. National Library of Medicine
National Institutes of Health

Search: PubMed
organelle organisation, cardiac muscle fiber development
Search Clear

Display Settings: Summary, 20 per page, Sorted by Recently Added
Send to: Filter your results: All (65) Review (4)

Results: 1 to 20 of 65

Excitation-contraction coupling changes during postnatal cardiac development.
Ziman AP, Gómez-Viquez NL, Bloch RJ, Lederer WJ.
J Mol Cell Cardiol. 2010 Feb;48(2):379-86. Epub 2009 Oct 8.
PMID: 19818794 [PubMed - indexed for MEDLINE]
[Related citations](#)

Bin1 SRC homology 3 domain acts as a scaffold for myofiber sarcomere assembly.
Fernando P, Sandoz JS, Ding W, de Repentigny Y, Brunette S, Kelly JF, Kothary R, Megre J.
J Biol Chem. 2009 Oct 2;284(40):27674-86. Epub 2009 Jul 26.
PMID: 19633357 [PubMed - indexed for MEDLINE]
[Related citations](#)

Wnt-11 signalling controls ventricular myocardium development by patterning N-cadherin.
Nagy II, Railo A, Rapila R, Hast T, Sormunen R, Tavi P, Räsänen J, Vainio SJ.
Cardiovasc Res. 2010 Jan 1;85(1):100-9. Epub .
PMID: 19622544 [PubMed - indexed for MEDLINE]
[Related citations](#)

An integrated strategy to study muscle development and myofilament structure in Caenorhabditis elegans.
Meissner B, Warner A, Wong K, Dube N, Lorch A, McKay SJ, Khattra J, Rogalski T, Sornmo L, Miller DM 3rd, Baillie DL, Holt RA, Jones SJ, Marra MA, Moerman DG.
PLoS Genet. 2009 Jun;5(6):e1000537. Epub 2009 Jun 26.
PMID: 19557190 [PubMed - indexed for MEDLINE] [Free PMC Article](#) [Free text](#)
[Related citations](#)

Obscurin interacts with a novel isoform of MyBP-C slow at the periphery of the sarcomeric filament assembly.
Ackermann MA, Hu LY, Bowman AL, Bloch RJ, Kontogianni-Konstantopoulos A.
Mol Biol Cell. 2009 Jun;20(12):2963-78. Epub 2009 Apr 29.
PMID: 19403693 [PubMed - indexed for MEDLINE] [Free PMC Article](#) [Free text](#)
[Related citations](#)

Boolean search + specialization

- + Extend the query
- Number of results
- Results are difficult to understand
 - Which concepts are taken into account?
 - Which ones have been added?
- No relevance assessment
 - ⇒ Loss of the first query context

Nature Precedings : doi:10.1038/npre.2010.5408.1 : Posted 17 Dec 2010

Context: information retrieval using ontologies

?

Organelle organization (GO_0006996)
Cardiac muscle fiber development (GO_0048739)

Nature Precedings : doi:10.1038/npre.2010.5408.1 · Posted 17 Dec 2010



HOME SEARCH SITE MAP PubMed All Databases Human Genome
Search across databases (Organelle organization [gene ontology]

- Result counts displayed in gray indicate one or more terms not found

7		PubMed: biomedical literature citations and abstracts	none	
41		PubMed Central: free, full text journal articles	none	
none		Site Search: NCBI web and FTP sites	none	
none		OMIA: online Mendelian Inheritance in Animals	none	

none		Nucleotide: Core subset of nucleotide sequence records	none		dbGaP: genotype and phenotype
none		EST: Expressed Sequence Tag records	none		UniGene: gene-oriented clusters of transcript sequences
none		GSS: Genome Survey Sequence records	none		CDD: conserved protein domain database
none		Protein: sequence database	none		3D Domains: domains from Entrez Structure
none		Genome: whole genome sequences	none		UniSTS: markers and mapping data
none		Structure: three-dimensional macromolecular structures	none		PopSet: population study data sets
none		Taxonomy: organisms in GenBank	8668		GEO Profiles: expression and molecular abundance profiles
none		SNP: single nucleotide polymorphism	none		GEO DataSets: experimental sets of GEO data
none		dbVar: Genomic structural variation	none		Cancer Chromosomes: cytogenetic databases
13		Gene: gene-centered information	none		PubChem BioAssay: bioactivity screens of chemical substances
none		SRA: Sequence Read Archive	none		PubChem Compound: unique small molecule chemical structures
none		BioSystems: Pathways and systems of interacting molecules	none		PubChem Substance: deposited chemical substance records

Archive! Ensembl

Home Login / Register | BLAST/BLAT | BioMart | Docs & FAQs | Mirrors

New Count Results URL XML Perl Help

Please restrict your query using criteria below

Dataset: Homo sapiens genes (GRCh37)

Filters: GO Biological Process Term: GO:0048739

Attributes: Ensembl Gene ID

Dataset: [None Selected]

REGION:

GENE:

TRANSCRIPT EVENT:

GENE ONTOLOGY: Evidence code (GO Biological process) GO Biological Process Term: GO:0048739

Number of retrieved genes

GenBank	0
Ensembl	0
AND	0
OR	13



Take better benefits of ontologies during the information retrieval process (indexing/query matching)

- Expand the query if necessary
- Measure document/query adequacy by identifying added concepts

Favor the overall results' grasp by the user

- Explain why a document has been selected
- Give an overall vision of results
- If a selected document is not relevant, identify why in order to reformulate the query conveniently

Taking user preferences into account

⇒ Favor interactions and iterative querying process

Context and objectives

Ontology based information retrieval

Relevance calculus between a document index and a query

Similarity between two concepts

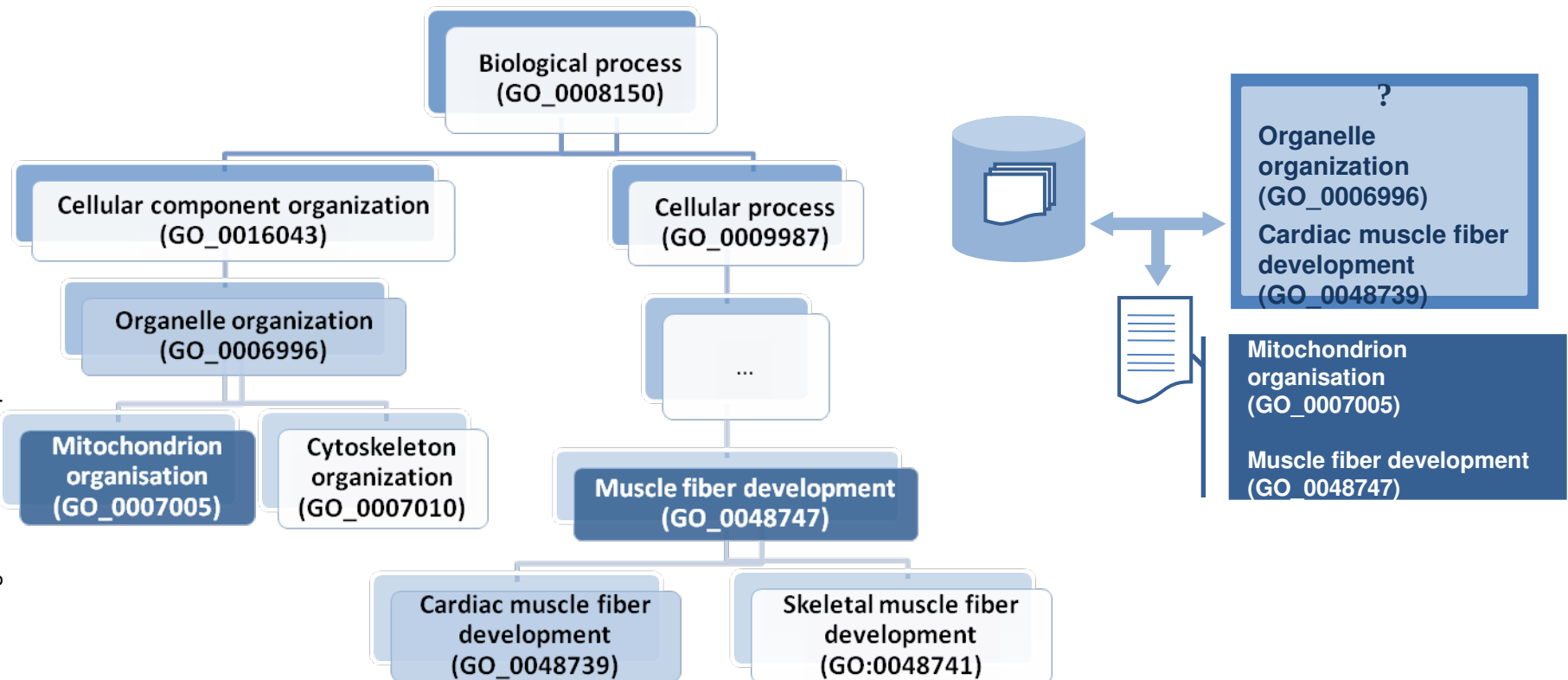
Relevance of a document with respect to a concept

Relevance of a document with respect to a query

Results visualization

Conclusion et perspectives

Hyponyms and hypernyms to avoid silences



- ⇒ Mix documents that match more or less the query
- ⇒ The selection may be difficult to understand

Context and objectives

Ontology based information retrieval

Relevance calculus between a document index and a query

Similarity between two concepts

Relevance of a document with respect to a concept

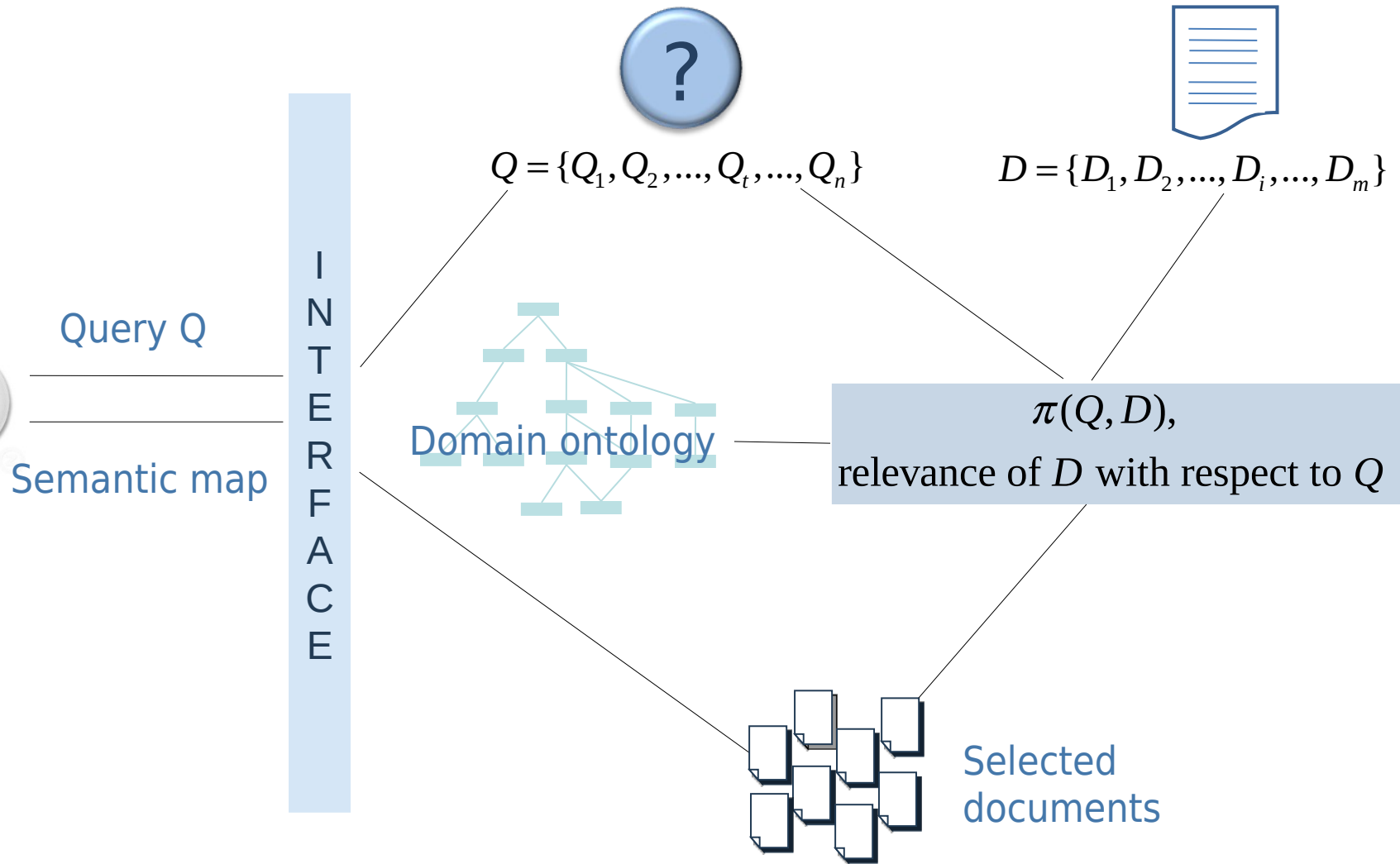
Relevance of a document with respect to a query

Results visualization

Conclusion et perspectives

Relevance calculus between a document index and a query

Nature Precedings : doi:10.1038/npre.2010.5408.1 : Posted 17 Dec 2010



Three-level relevance calculus

- Similarity between two concepts: a concept from the document index and a concept from the query $\pi(Q_t, D_i)$
- Relevance of a document (i.e. the set of its indexing concepts) with respect to a concept from the query $\pi(Q_t, D)$
- Relevance of a document with respect to a query: Fuzzy aggregation of relevance measures $\pi(Q, D)$

⇒ Advantages

- Ranking of documents with respect to their relevance
- Detailed explanation of the document selection

Three-level relevance calculus

- Similarity between two concepts: a concept from the document index and a concept from the query $\pi(Q_t, D_i)$
 - Relevance of a document (i.e. the set of its indexing concepts) with respect to a concept from the query $\pi(Q_t, D)$
 - Relevance of a document with respect to a query: Fuzzy aggregation of relevance measures $\pi(Q, D)$
- Several similarity measurements have been proposed in literature, this one is easy to understand (% of mutual hyponyms)

$$\pi_{JD}(C_1, C_2) = \frac{\text{hypo}(C_1) \cap \text{hypo}(C_2)}{\text{hypo}(C_1) \cup \text{hypo}(C_2)}, \text{ if } C_1 \text{ } \blacklozenge \text{ hypo}(C_2) \text{ ou } C_2 \text{ } \blacklozenge \text{ hypo}(C_1)$$

0, else

Relevance calculus between a document index and a query

Three-level relevance calculus

- Similarity between two concepts: a concept from the document index and a concept from the query $\pi(Q_t, D_i)$
- Relevance of a document (i.e. the set of its indexing concepts) with respect to a concept from the query $\pi(Q_t, D)$
- Relevance of a document with respect to a query: Fuzzy aggregation of relevance measures $\pi(Q, D)$

- Best relevance between indexing concepts of document D and a query concept Q_t

$$\pi(Q_t, D) = \max_{i \in |D|} \pi(Q_t, D_i)$$

- May be generalized by weighting the concepts D_i (using *evidence codes* in the Gene Ontology for example)

Relevance calculus between a document index and a query

Three-level relevance calculus

- Similarity between two concepts: a concept from the document index and a concept from the query $\pi(Q_t, D_i)$
- Relevance of a document (i.e. the set of its indexing concepts) with respect to a concept from the query $\pi(Q_t, D)$
- Relevance of a document with respect to a query: Fuzzy aggregation of relevance measures $\pi(Q, D)$
 - Combine individual relevance scores to estimate an overall relevance of the document
 - Take user preferences into account: decision theory
 - Yager operator (with $q \in \mathbf{R}$)

$$Y_m(\pi(Q_1, D), \dots, \pi(Q_{|Q|}, D)) = \prod_{i=1}^{|Q|} \pi(Q_t, D)^{q/|Q|}$$

- $q = 1$, arithmetic mean,
- $q = -1$, harmonic mean,
- $q \rightarrow 0$, geometrical mean,
- $q \rightarrow +\infty$, max (OR generalization)
- $q \rightarrow -\infty$, min (AND generalization)

Context and objectives

Ontology based information retrieval

Relevance calculus between a document index and a query

Similarity between two concepts

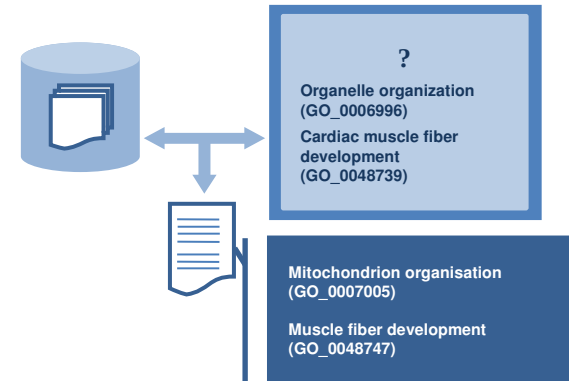
Relevance of a document with respect to a concept

Relevance of a document with respect to a query

Results visualization

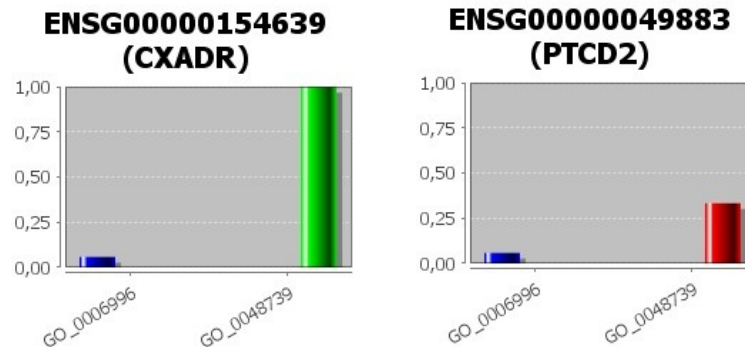
Conclusion et perspectives

- A document may be selected even if its index contains no terms of the query



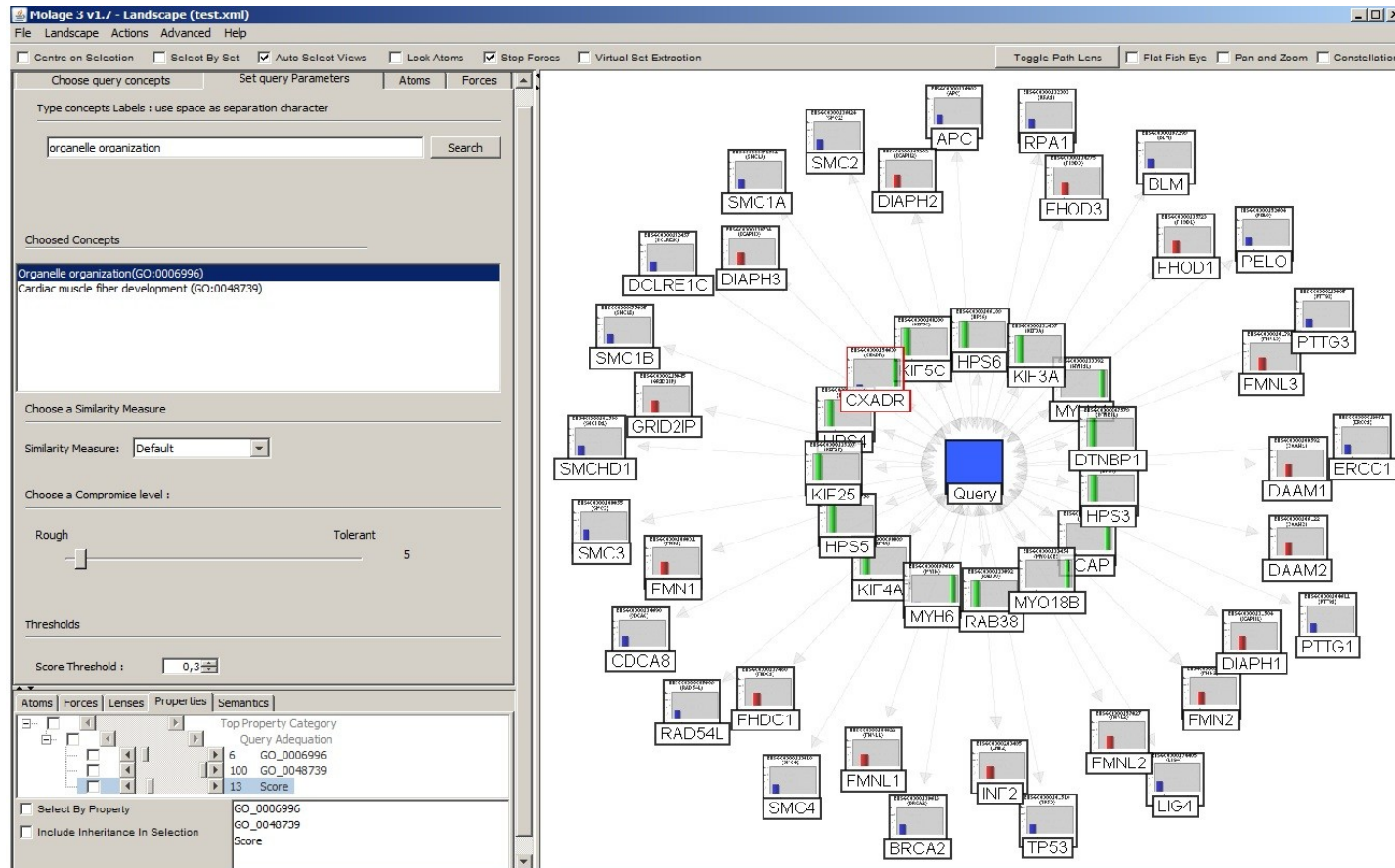
Explain the selection to the user: pictograms

- Each concept of the query is associated with a bar:
 - Its height is proportional to its relevance $\pi(Q_t, D)$
 - Its color says if Q_t
 - index the document (D)
 - specialize (is an hyponym of) D_i
 - generalize (is an hypernym of) D_i



Pictograms are displayed on a semantic map

- Their physical distance to the query is proportional to their relevance score: $\pi(Q, D)$
- Visualization and navigation: fit the human cognitive limits (lens, number of results, relevance threshold...) and help the user (selection of concept for the query...)



Context and objectives

Ontology based information retrieval

Relevance calculus between a document index and a query

Similarity between two concepts

Relevance of a document with respect to a concept

Relevance of a document with respect to a query

Results visualization

Conclusion et perspectives

Results

- Find **more documents** (avoid silences)
- Improve the **relevance**: documents **ranking**
- **Explain** relevance calculus (diagnose)
- **Visualize** overall results
- **Interaction** with the list of retrieved documents: customize user preferences
 - ⇒ Iterative improvement of the query

Perspectives

- Improve CHI
- Suggest **query reformulation**
 - From documents selection by the user (weighting + complement)
 - Underline query terms that are discriminated
- Test several **semantic distance calculus** on different benchmarks (TREC, Much more...)
- Improve visualization
 - Filter the displayed results using **sub-ontologies extraction**
 - Propose a view of the results underlining **clusters**
- Propose an online version

User Centered and Ontology Based Information Retrieval System for Life Sciences



sylvie.ranwez@mines-ales.fr
vincent.ranwez@univ-montp2.fr
mohameth.sy@mines-ales.fr
jacky.montmain@mines-ales.fr
michel.crampes@mines-ales.fr

OBIRS on line: <http://www.ontotoolkit.mines-ales.fr/ObirsClient/>

Nature Precedings : doi:10.1038/npre.2010.5408.1 : Posted 17 Dec 2010

OBIRS: Query Concepts Results

Parameters

Choose a Similarity Measure

Choose an Agregator Operator

Choose a Compromise level
 Rough Tolerant

Thresholds
 Score Threshold:
 Number of results:

View

PTCD2

Category	Score
0006996	~0.05
0048739	~0.35

Match Explanation

organelle organization : score 6 (obtain by generalisation of mitochondrion organization)

cardiac muscle fiber development : score 33 (obtain by specialisation of muscle fiber development)

Score: 70

21

User Centered and Ontology Based Information Retrieval System for Life Sciences – S . Ranwez