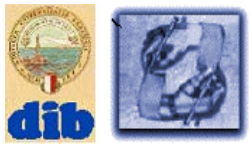
DELOS
Network of Excellence on Digital Libraries

DELOS All Tasks Meeting

Paris, France, 30-31 January 2006

# ITem Recommender (ITR)

## A Content-based Recommender for Personalized Information Access

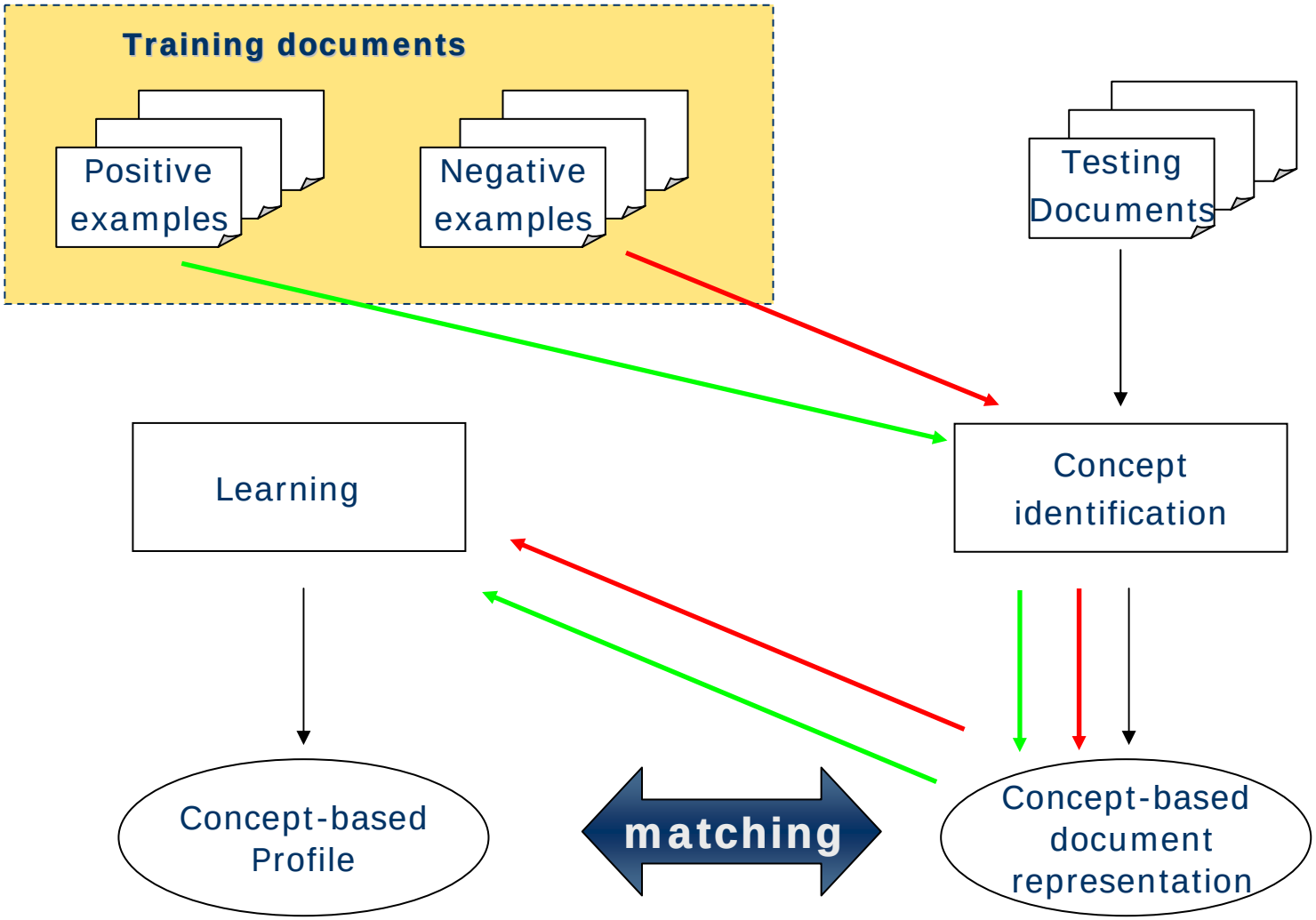



Giovanni Semeraro  Pasquale Lops

Marco Degemmis  Pierpaolo Basile

*LACAM – Knowledge Acquisition and Machine Larning Lab
Department of Informatics – University of Bari

# Intelligent Personalized Information Access



# Application Domain: Paper Recommending

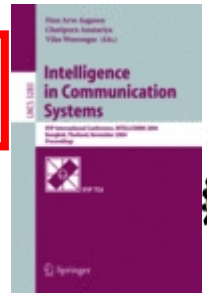
**Reducing OWL Entailment to Description Logic Satisfiability**

Ian Horrocks[1] and Peter F. Patel-Schneider[2]

(1) Department of Computer Science, University of Manchester,

(2) Bell Labs Research, Lucent Technologies,

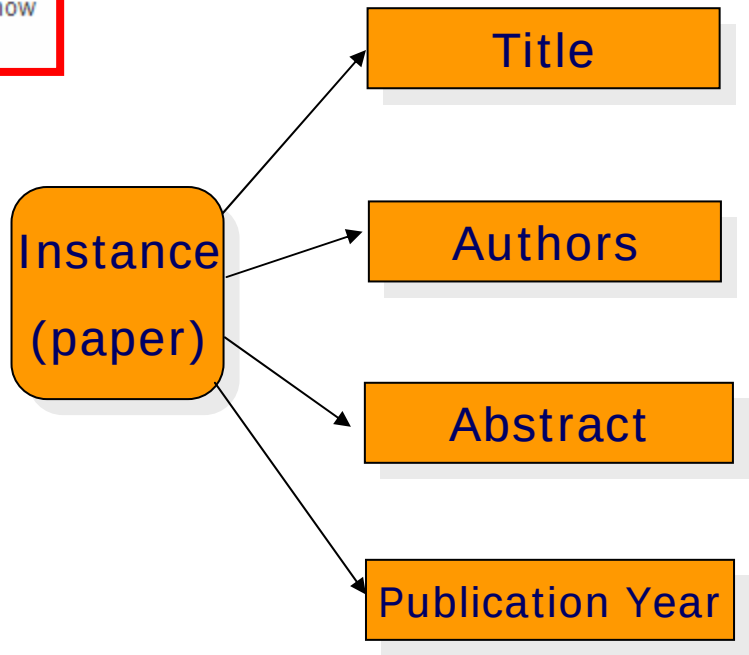
We show how to reduce ontology entailment for the OWL DL and OWL Lite ontology languages to knowledge base satisfiability in (respectively) the (**D**) and (**D**) description logics. This is done by first establishing a correspondence between OWL ontologies and description logic knowledge bases and then by showing how knowledge base entailment can be reduced to knowledge base satisfiability.




Tokenization + Stopword + Stemming → **Keyword-based representation**

Tokenization + Stopword + POS + disambiguation → **Sense-based representation**


User Ratings



# Bag of Synsets

## Bag of Words

| *Id paper* | *Word Form* | *Occurrence* |
|---|---|---|
| 31 | artificial | 2 |
| 31 | intelligence | 2 |
| … | … | … |
| 1134 | WWW | 3 |
| 1134 | web | 2 |
| … | … | … |

## Bag of Synsets

| *Id paper* | *Word Form* | *Id Synset* | *Occurrence* |
|---|---|---|---|
| 31 | artificial intelligence | 6712568 | 2 |
| … | … | | … |
| 1134 | WWW,web | 04425517 | 5 |
| … | … | | … |

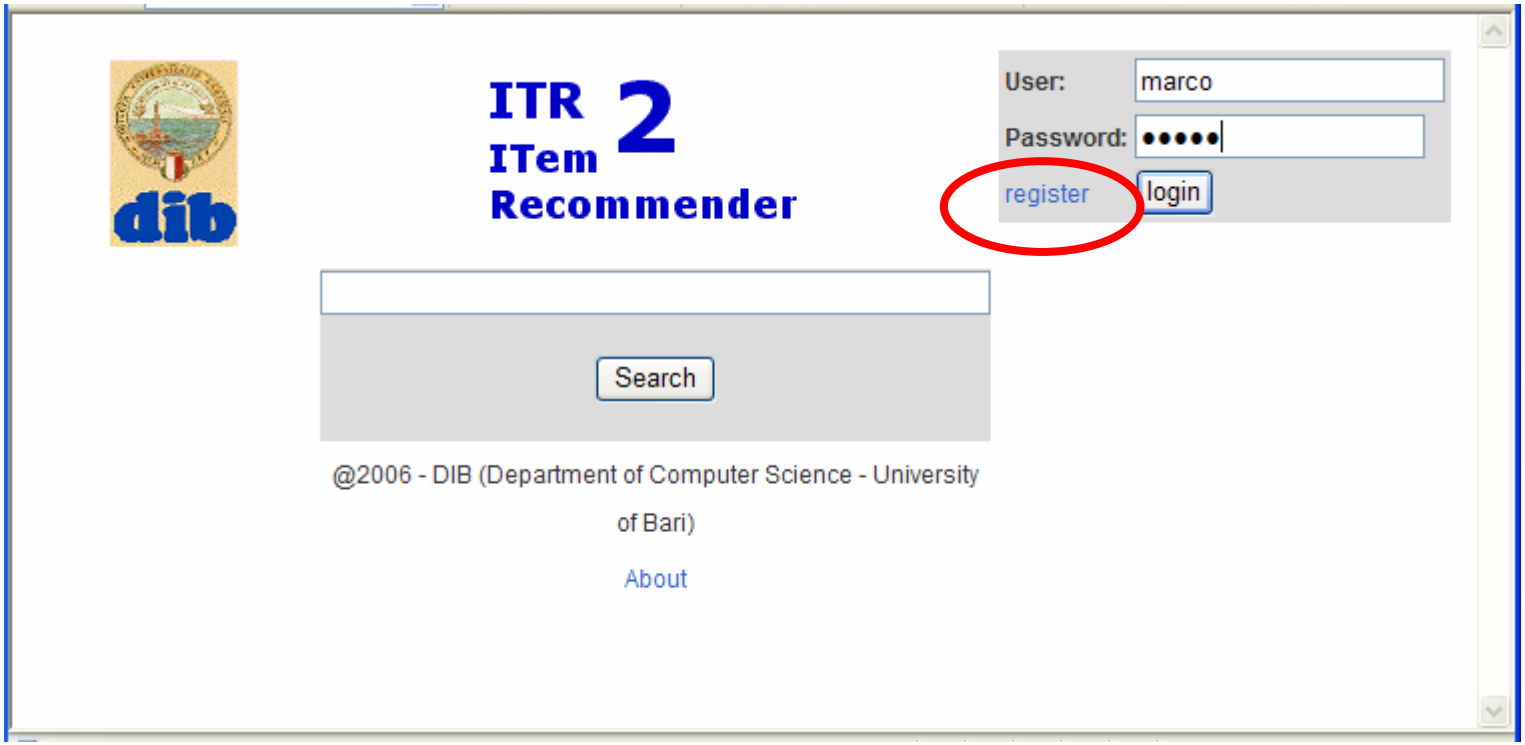
- Reduction of features representing papers in the dataset
  - ✓ Mainly on slots containing proper names
  - ✓ Recognition of bigrams
  - ✓ Synonyms represented by the same synsets

# Naïve Bayes Classification: ITem Recommender

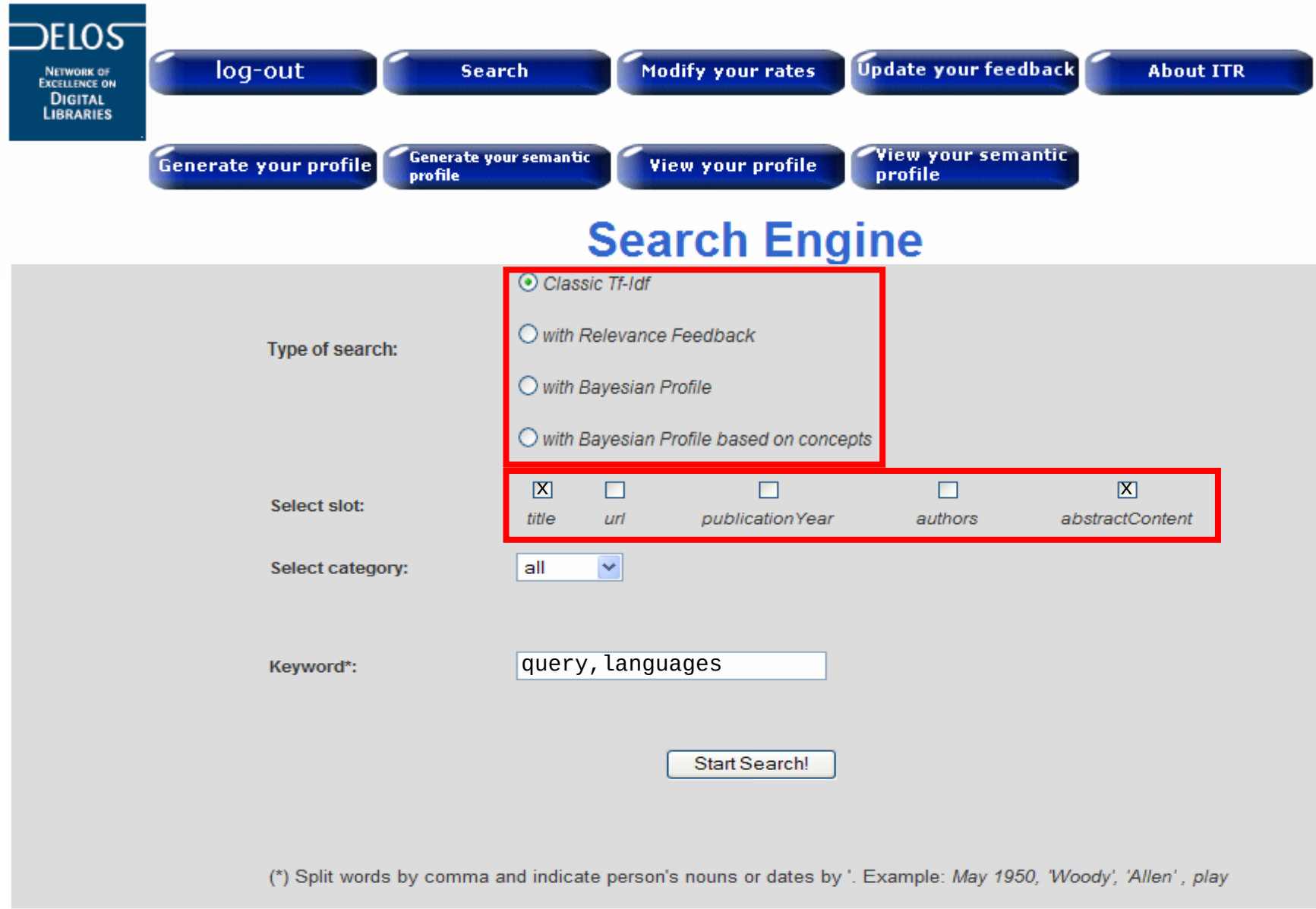
## ITem Recommender (ITR)

1. Content-based item recommending on the basis of ratings given by users
2. Naïve Bayes text classification to assign a score (level of interest) to items according to the user preferences
   - ✓ Performance comparable to more complex algorithms
   - ✓ Increasingly popular in text classification
3. Result: user profile containing the probabilistic model of user preferences

# Home page


ITR 2
ITem
Recommender
User:
marco
Password:
register
login
Search
@2006 - DIB (Department of Computer Science - University
of Bari)
About

# Personalized Search process



# Keyword-based Query

**Type of search:** classic

**Selected slot** title abstractContent

**Selected category:** all

**Keywords:** query, languages

Search

***Maybe you find "query" with sense:***

***Noun:***

- *question, inquiry, enquiry, query, interrogation -- (an instance of questioning; "there was a question about my training"; "we made inquiries of all those who were present")], [5998352]*

***Verb:***

- *question, query -- (pose a question)], [625049]*

***Maybe you find "languages" with sense:***

# Keyword-based search results

# Keyword-based search results: Top ranked item

**Genre:** dummy

**title:** Semantic Processing of the Semantic Web

**url:** http://www.springerlink.com/openurl.asp?genre=article&id=0PKNJGTVJBNWPUNT

**publicationYear:** 2003

**authors:** Patel, Kunal Gupta, Gopal

**abstractContent:** We develop a semantics based approach to process information on the semantic web. We show how Horn logic can be used to denotationally capture the semantics of mark up **languages** designed for describing resources on the semantic web (such as RDF). The same approach can also be used to specify the semantics of **query languages** for the semantic web. The semantics of both the resource description **languages** and the **query languages** are executable and when put together can be used to compute answers to semantic web **queries.** The main advantage of this semantic based approach to processing the semantic web is that these executable semantics can be developed extremely quickly. Thus, as the semantic web mark up **languages** evolve rapidly, their implementations can be developed at the same pace. In this paper, we present our approach based on denotational semantics and Horn logic. Our approach is quite general, and applicable to any description format (XML, RDF, DAML, etc.), though in this paper we illustrate it via RDF (Resource Description Framework).

Rates: ○ No rate ○ 1.0 ○ 2.0 ○ 3.0 ○ 4.0 ○ 5.0 ○ 6.0 ◉ 7.0 ○ 8.0 ○ 9.0 ○ 10.0 Vote

# Sense-based Query

***Maybe you find "languages" with sense:***

***Noun:***

- [ ] *language, linguistic_communication -- (a systematic means of communicating by the use of sounds or conventional symbols; "he taught foreign languages"; "the language introduced is standard throughout the text"; "the speed with which a program can be executed depends on the language in which it is written")], [5294998]*

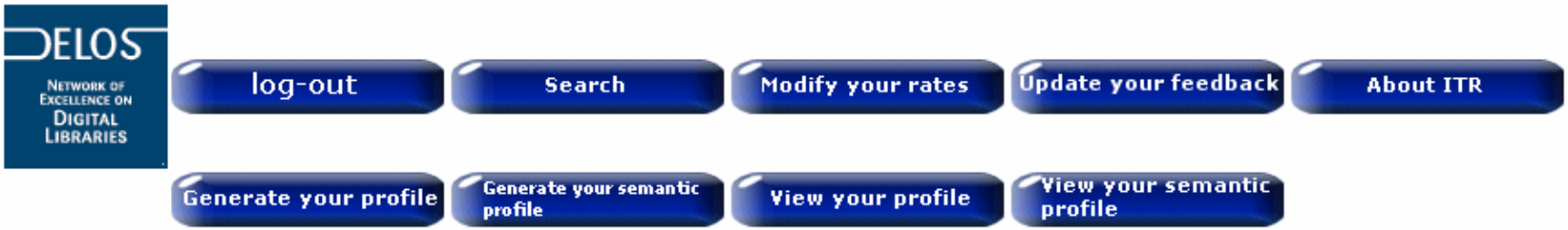
- [ ] *speech, speech_communication, spoken_communication, spoken_language, language, voice_communication, oral_communication -- (communication by word of mouth; "his speech was slurred"; "the telephone greatly increased the range of speech communication"; "he uttered harsh language"; "he recorded the spoken language of the streets")], [5933245]*

- [x] *terminology, nomenclature, language -- (a system of words used in a particular discipline; "legal terminology"; "the language of sociology")], [5309228]*

- [ ] *linguistic_process, language -- (the cognitive processes involved in producing and understanding linguistic communication; "he didn't have the language to express his feelings")], [4977017]*

- [ ] *language, speech -- (the mental faculty or power of vocal communication; "language sets homo sapiens apart from all other animals")], [4852301]*

# Sense-based search results




## Ranking method : Classic TF-IDF

### Items found: 4

**Genre:** dummy

**title:** Knowledge-Intensive Induction of Terminologies from Metadata

**url:** http://www.springerlink.com/openurl.asp?genre=article&id=29JBEFRCWG1D2B8C

**publicationYear:** 2004

**authors:** Esposito, Floriana Fanizzi, Nicola Iannone, Luigi Palmisano, Ignazio Semeraro, Giovanni

**abstractContent:** We focus on the induction and revision of terminologies from metadata. Following a Machine Learning approach, this setting can be cast as a search pro...

Relevance : 0.36910153338033663 Show

sense cosine similarity

**Genre:** dummy

**title:** A Comparison of RDF Query Languages

**url:** http://www.springerlink.com/openurl.asp?genre=article&id=FTXY71QEDRHB945V

**publicationYear:** 2004

# Sense-based search results: Top ranked item

**Genre:** dummy

**title:** Knowledge Intensive Induction of **Terminologies** from Metadata

**url:** http://www.springerlink.com/openurl.asp?genre=article&id=29JBEFRCWG1D2B8C

**publicationYear:** 2004

**authors:** Esposito, Floriana Fanizzi, Nicola Iannone, Luigi Palmisano, Ignazio Semeraro, Giovanni

**abstractContent:** We focus on the induction and revision of **terminologies** from metadata. Following a Machine Learning approach, this setting can be cast as a search problem to be solved employing operators that traverse the search space expressed in a structural representation, aiming at correct concept definitions. The progressive refinement of such definitions in a **terminology** is driven by the available extensional knowledge (metadata). A knowledge intensive inductive approach to this task is presented, that can deal with on the expressive Semantic Web representations based on Description Logics, which are endowed with well founded reasoning capabilities. The core inferential mechanism, based on multilevel counterfactuals, can be used for either inducing new concept descriptions or refining existing (incorrect) ones. The soundness of the approach and its applicability are also proved and discussed1.

Rates: ◉ No rate ○ 1.0 ○ 2.0 ○ 3.0 ○ 4.0 ○ 5.0 ○ 6.0 ○ 7.0 ○ 8.0 ○ 9.0 ○ 10.0 Vote

return to search results

# Example of Keyword-based User Profile

| User ID: 6 | Category: dummy | Class Priors: | P(YES)= 0.5333333 | P(NO)= 0.4666666 |
|---|---|---|---|---|

Slot: **abstractContent**

| Feature | Strength |
|---|---|
| entiti | 3.6138782 |
| repositori | 2.8947555 |
| easi | 2.8947555 |
| reusabl | 2.8947555 |
| upper-level | 2.8947555 |
| way | 2.8947555 |
| avail | 2.8947555 |
| area | 2.5893739 |
| kim | 2.4626222 |
| seamless | 2.1475411 |
| subtasks: | 2.1475411 |

$$strength(t_k, s_m) = \log \frac{P(t_k \mid c_+, s_m)}{P(t_k \mid c_-, s_m)}$$

Features are keywords

# Example of Sense-based User Profile

| User ID: 6 | Category: dummy | Class Priors: | P(YES)= 0.5333333 | P(NO)= 0.4666666 |
|---|---|---|---|---|

Slot: **abstractContent**

| Feature | Strength |
|---|---|
| 1742 | 3.6145387 |
| 2268652 | 2.8954161 |
| 2766412 | 2.8954161 |
| 2223910 | 2.8954161 |
| 4415376 | 2.8954161 |
| 5655492 | 2.5376664 |
| 5552847 | 2.3181007 |
| 2311478 | 2.1482017 |
| 5309075 | 2.1482017 |
| 1636312 | 2.1482017 |

$$strength(t_k, s_m) = \log \frac{P(t_k \mid c_+, s_m)}{P(t_k \mid c_-, s_m)}$$

is computed on synsets instead of keywords

Features are WordNet synsets

# Search results with keyword-based profile

## Ranking method : Bayes Profile

**Items found: 62**

**Genre:** dummy
**title:** Benchmarking DAML+OIL Repositories
**url:** http://www.springerlink.com/openurl.asp?genre=article&id=R0NQRF92QQDT7271
**publicationYear:** 2003
**authors:** Guo, Yuanbo Heflin, Jeff Pan, Zhengxiang
**abstractContent:** We present a benchmark that facilitates the evaluation of DAML+OIL repositories in a standard and systematic way. This benchmark is intended to evalua...
Relevance : 0.05014214319292595 Show

computed by keyword-based user profile

## Ranking method : Classic TF-IDF

**Items found: 62**

**Genre:** dummy
**title:** Semantic Processing of the Semantic Web
**url:** http://www.springerlink.com/openurl.asp?genre=article&id=0PKNJGTVJBNWPUNT
**publicationYear:** 2003
**authors:** Patel, Kunal Gupta, Gopal
**abstractContent:** We develop a semantics based approach to process information on the semantic web. We show how Horn logic can be used to denotationally capture the sem...
Relevance : 0.38646091250299275 Show

word cosine similarity

# Search results with sense-based profile

## Ranking method : Profiles_semantic

**Items found: 4**

**Genre:** dummy
**title:** Knowledge-Intensive Induction of Terminologies from Metadata
**url:** http://www.springerlink.com/openurl.asp?genre=article&id=29JBEFRCWG1D2B8C
**publicationYear:** 2004
**authors:** Esposito, Floriana Fanizzi, Nicola Iannone, Luigi Palmisano, Ignazio Semeraro, Giovanni
**abstractContent:** We focus on the induction and revision of terminologies from metadata. Following a Machine Learning approach, this setting can be cast as a search pro...
Relevance : 0.17801534661207868 Show

computed by sense-based user profile

## Ranking method : Classic TF-IDF

**Items found: 4**

**Genre:** dummy
**title:** Knowledge-Intensive Induction of Terminologies from Metadata
**url:** http://www.springerlink.com/openurl.asp?genre=article&id=29JBEFRCWG1D2B8C
**publicationYear:** 2004
**authors:** Esposito, Floriana Fanizzi, Nicola Iannone, Luigi Palmisano, Ignazio Semeraro, Giovanni
**abstractContent:** We focus on the induction and revision of terminologies from metadata. Following a Machine Learning approach, this setting can be cast as a search pro...
Relevance : 0.36910153338033663 Show

sense cosine similarity

# Future Work

1. Ontology-based user context models
2. Integration of specific domain ontologies in the WSD algorithm
3. Representation of user models in RDF language

# Contacts

Link:

http://193.204.187.223:8080/EMIRDELOS/


| | |
|---|---|
| G. Semeraro | semeraro@di.uniba.it |
| P. Lops | lops@di.uniba.it |
| M. Degemmis | degemmis@di.uniba.it |
| P. Basile | basilepp@di.uniba.it |

# Classification Phase

1. Each document is represented as a vector of BOS, one for each slot
2. Each slot is independent from the others

$$P(c_j \mid d_i) = \frac{P(c_j)}{P(d_i)} \prod_{m=1}^{|S|} \prod_{k=1}^{|b_{im}|} P(t_k \mid c_j, s_m)^{n_{kim}}$$

$S = \{s_1, s_2, \ldots, s_{|S|}\}$ is the set of slots

$b_{im}$ is the BOS in slot $s_m$ of instance $d_i$

$t_k$ is the $k^{th}$ token (occurring $n_{kim}$ times in BOS $b_{im}$)

# Semantic Indexing Procedure

A document *d* mapped into a list of WordNet synsets in 3 steps:

1. Each monosemous unigram *w* in a slot of the document *d* mapped into the corresponding WordNet synset;
2. For each *bigram* $<w_1, w_2>$ of the type $<noun, noun>$ or $<adjective, noun>$, a search in WordNet for one synset is performed. If it exists, WSD algorithm applied to the bigram, otherwise applied separately to $w_1$ and $w_2$, using all words in the slot as the context *C* of *w*;
3. Each polysemous unigram *w* disambiguated using all words in the slot as the context *C* of *w*.