CONICET AGENCIA

Learning Better Context Characterizations: an Intelligent Information Retrieval Approach

Carlos M Lorenzetti

cml@cs.uns.edu.ar

Ana G Maguitman

agm@cs.uns.edu.ar

Grupo de Investigación en Recuperación de Información y Gestión del Conocimiento

Laboratorio de Investigación y Desarrollo en Inteligencia Artificial

Universidad Nacional del Sur
Av. L.N. Alem 1253
Bahía Blanca - Argentina

Information Retrieval limitations

Information Retrieval limitations

Java ¿Es útil? Sí | No

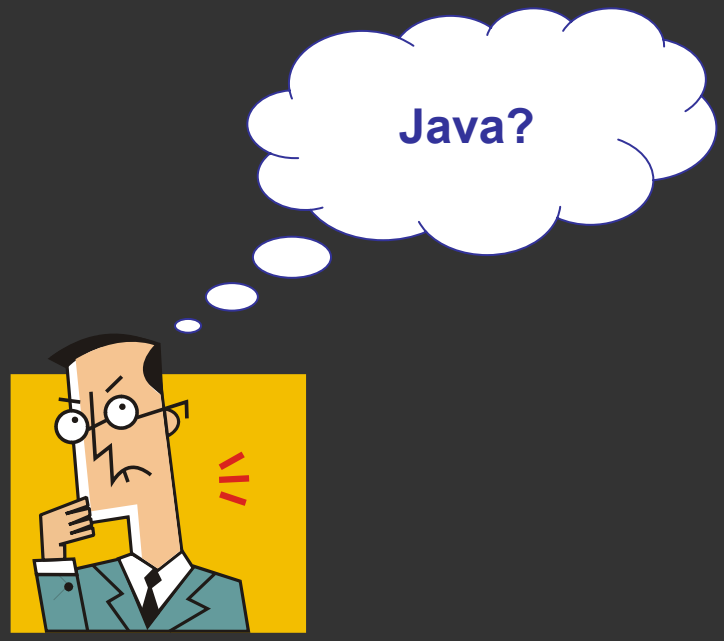
Isla del archipiélago Malayo situada en el sur de Indonesia. Al norte limita con el mar de Java, al este con el estrecho de Bali, al sur con el océano Índico, y al oeste con el estrecho de la Sonda. Yakarta es la ciudad más grande y la capital del país.

Enciclopedia Encarta

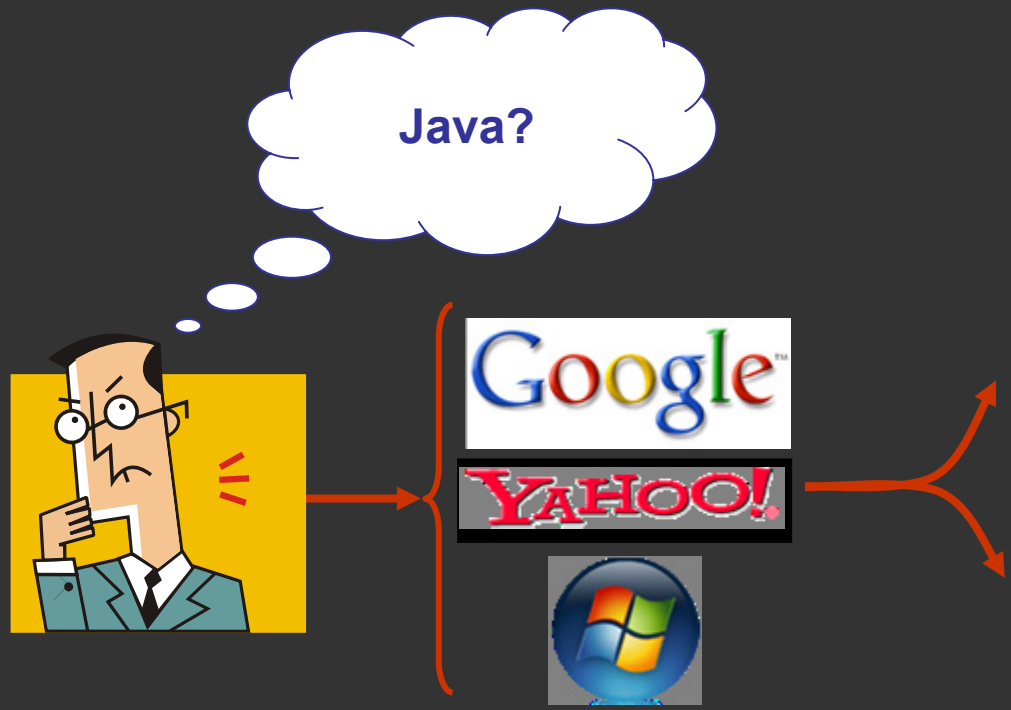
Java as an island

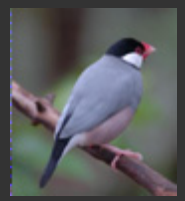
Information Retrieval limitations

Problems: ambiguity



Problems: ambiguity





Animals

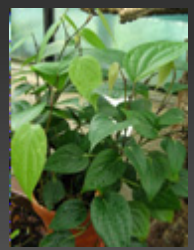
Computers

Entertainment

British American Tobacco

Consumables

Geography



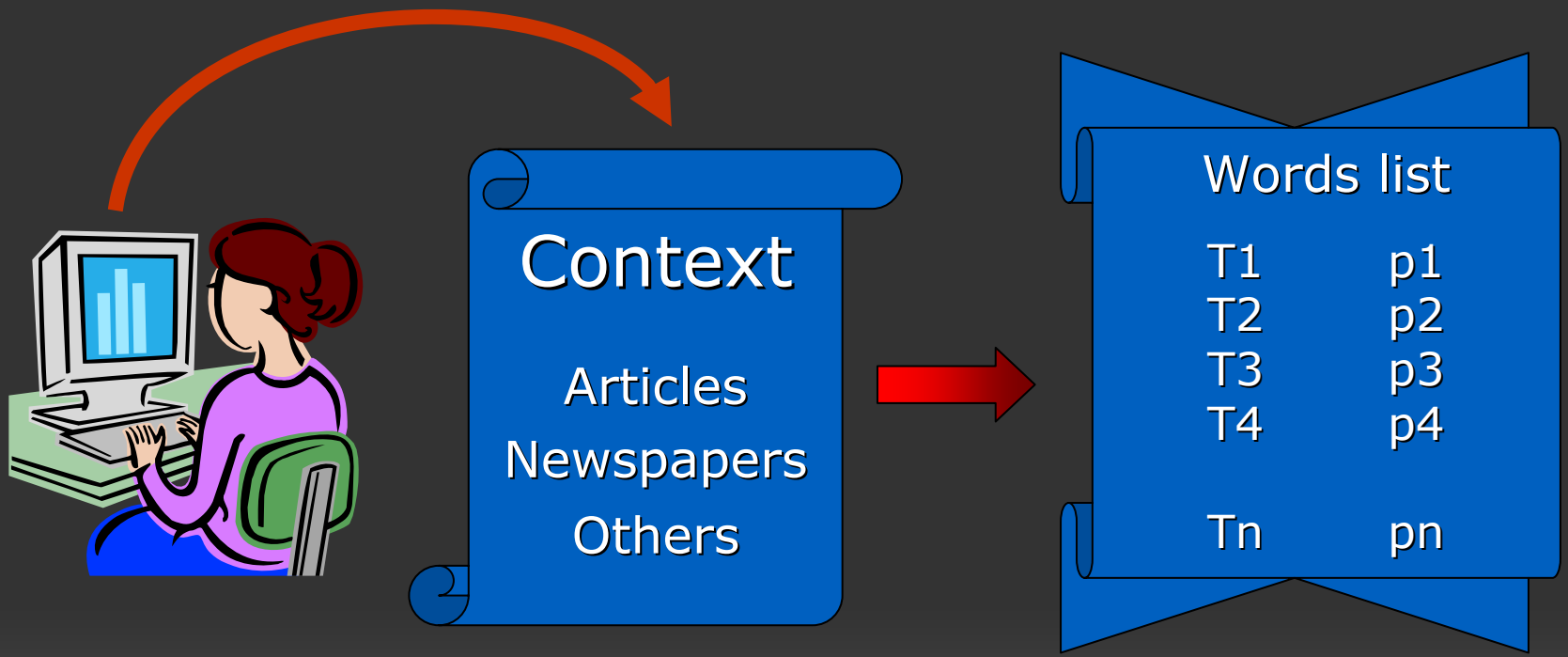
Flora

Ships

Proposed solutions

- Context-Specific term identification.
- Find relevant information sources.
- Incrementally make queries.

Context Characterization



Context Characterization

- Terms relevance
- TF-IDF

it uses the most simple way

$$TFIDF(d,t) = TF(d,t) \times IDF(t)$$

Counts documents' term ocurrence

Penalizes very common terms

Different Role of Terms

- *Descriptors*

  Terms that appear often in documents related to the given topic

  What is this topic about?

- Discriminators

  Terms that appear only in documents related to the given topic

  What terms are useful to seek similar information?

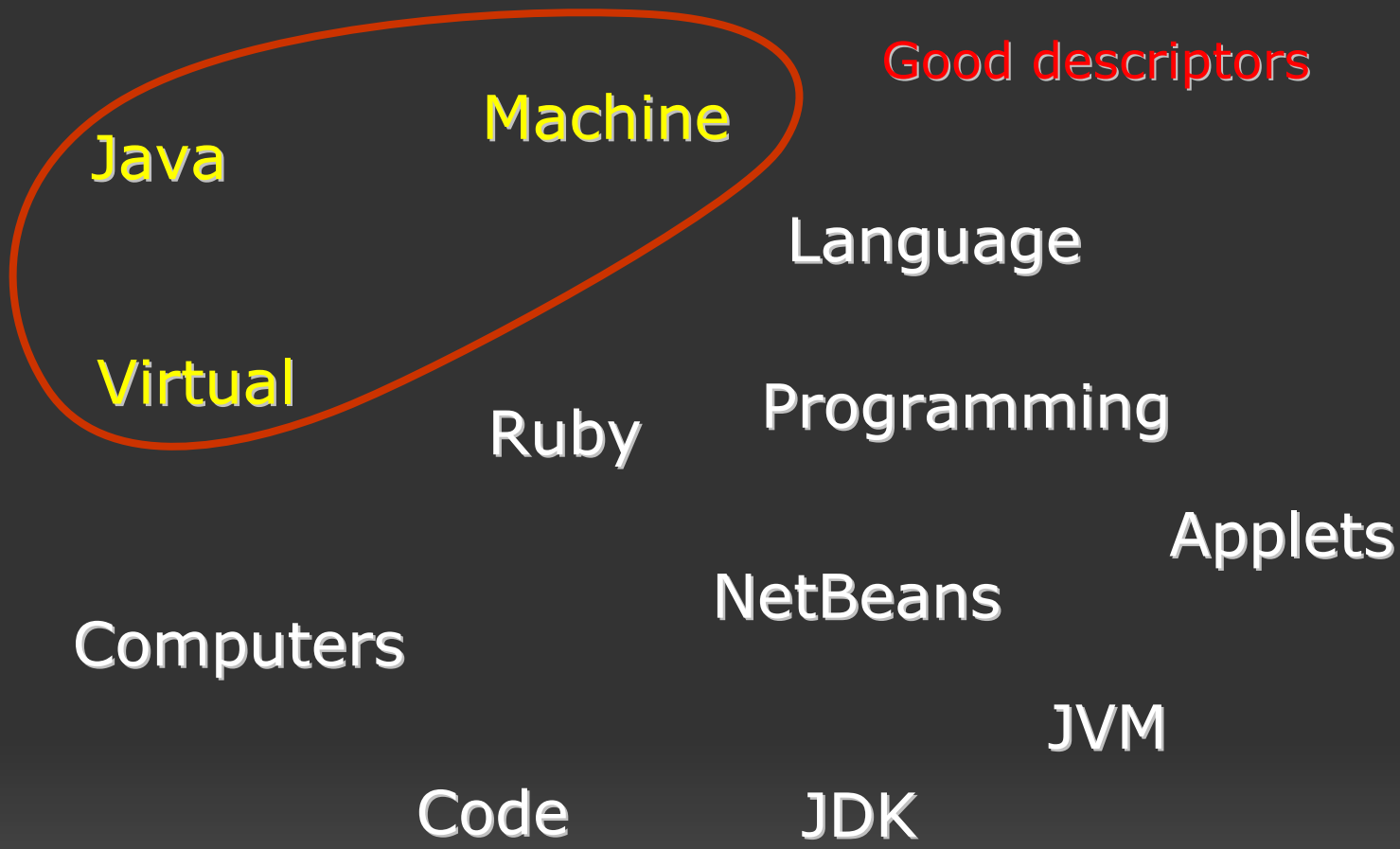
Descriptors and Discriminators Computation: an example

Descriptors and Discriminators

Java
Machine
Language
Virtual
Ruby
Programming
Applets
NetBeans
Computers
JVM
Code
JDK

Topic: Java Virtual Machine

Descriptors and Discriminators



Topic: Java Virtual Machine

Descriptors and Discriminators

Java

Machine

Language

Virtual

Ruby

Programming

Applets

NetBeans

Computers

JVM

Code

JDK

Good discriminators

Topic: Java Virtual Machine

Documents Descriptors and Discriminators

Initial Context		H (1)	(2)	(3)	(4)
java	4	2	5	5	2
machine	2	6	3	2	0
virtual	1	0	1	1	0
language	1	0	2	1	1
programming	3	0	2	2	0
coffee	0	3	0	0	3
island	0	4	0	0	2
province	0	4	0	0	1
jvm	0	0	2	1	0
jdk	0	0	3	3	0

Topic: Java Virtual Machine

(1) espressotec.com
(2) netbeans.org
(3) sun.com
(4) wikitravel.org

$$\mathbf{H}[d_i, t_j] = k$$

Number of occurrences of term *j* in document *i*

Documents *Descriptors*

Initial Context		$\lambda(d_0, t_j)$
java	4	0,718
machine	2	0,359
virtual	1	0,180
language	1	0,180
programming	3	0,539
coffee	0	0,000
island	0	0,000
province	0	0,000
jvm	0	0,000
jdk	0	0,000

Topic: Java Virtual Machine

Descriptive power of a term in a document

$$\lambda(d_i, t_j) = \frac{\mathbf{H}[i, j]}{\sqrt{\sum_{k=0}^{n-1} (\mathbf{H}[i, k])^2}}$$

Documents *Discriminators*

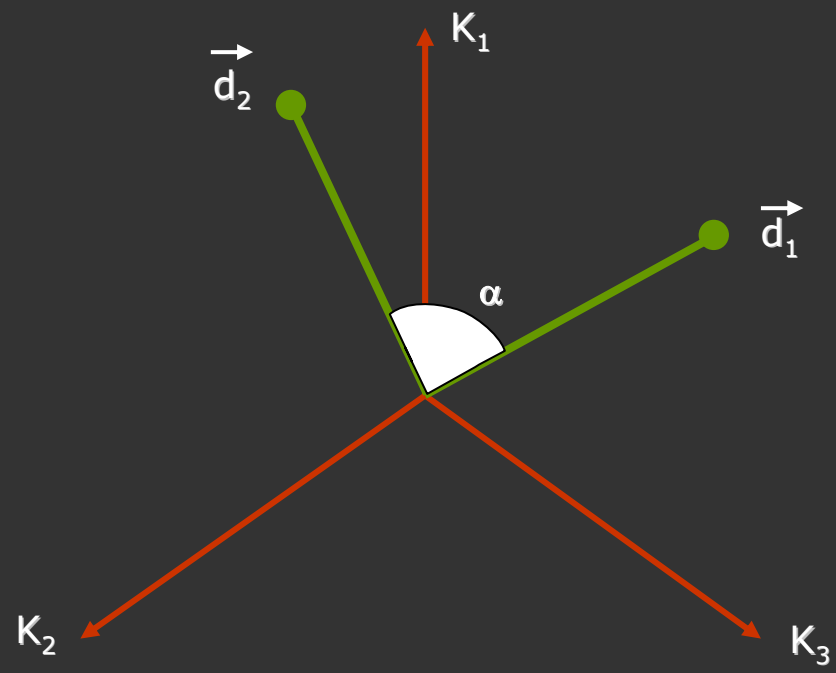
Initial Context		$\delta(t_i, d_0)$
java	4	0,447
machine	2	0,500
virtual	1	0,577
language	1	0,500
programming	3	0,577
coffee	0	0,000
island	0	0,000
province	0	0,000
jvm	0	0,000
jdk	0	0,000

Topic: Java Virtual Machine

Discriminating power of a term in a document

$$\delta(t_i, d_j) = \frac{s(\mathbf{H}^T[i,j])}{\sqrt{\sum_{k=0}^{m-1} s(\mathbf{H}^T[i,k])}}$$

Documents comparison criteria



Cosine similarity

$$\sigma(d_i, d_j) = \sum_{k=0}^{n-1} (\lambda(d_i, t_k) \cdot \lambda(d_j, t_k))$$

Documents similarity

Topics Descriptors

Initial Context		$\Lambda(d_0, t_j)$
java	4	0,385
machine	2	0,158
virtual	1	0,124
language	1	0,089
programming	3	0,064
coffee	0	0,055
island	0	0,040
province	0	0,040
jvm	0	0,032
jdk	0	0,014

Topic: Java Virtual Machine

Term descriptive power in a topic of a document

$$\Lambda(d_i, t_j) = \frac{\sum_{k=0,\ k\neq i}^{m-1} (\sigma(d_i, d_k) \cdot \lambda(d_k, t_j)^2)}{\sum_{k=0,\ k\neq i}^{m-1} \sigma(d_i, d_k)}$$

Topics Discriminators

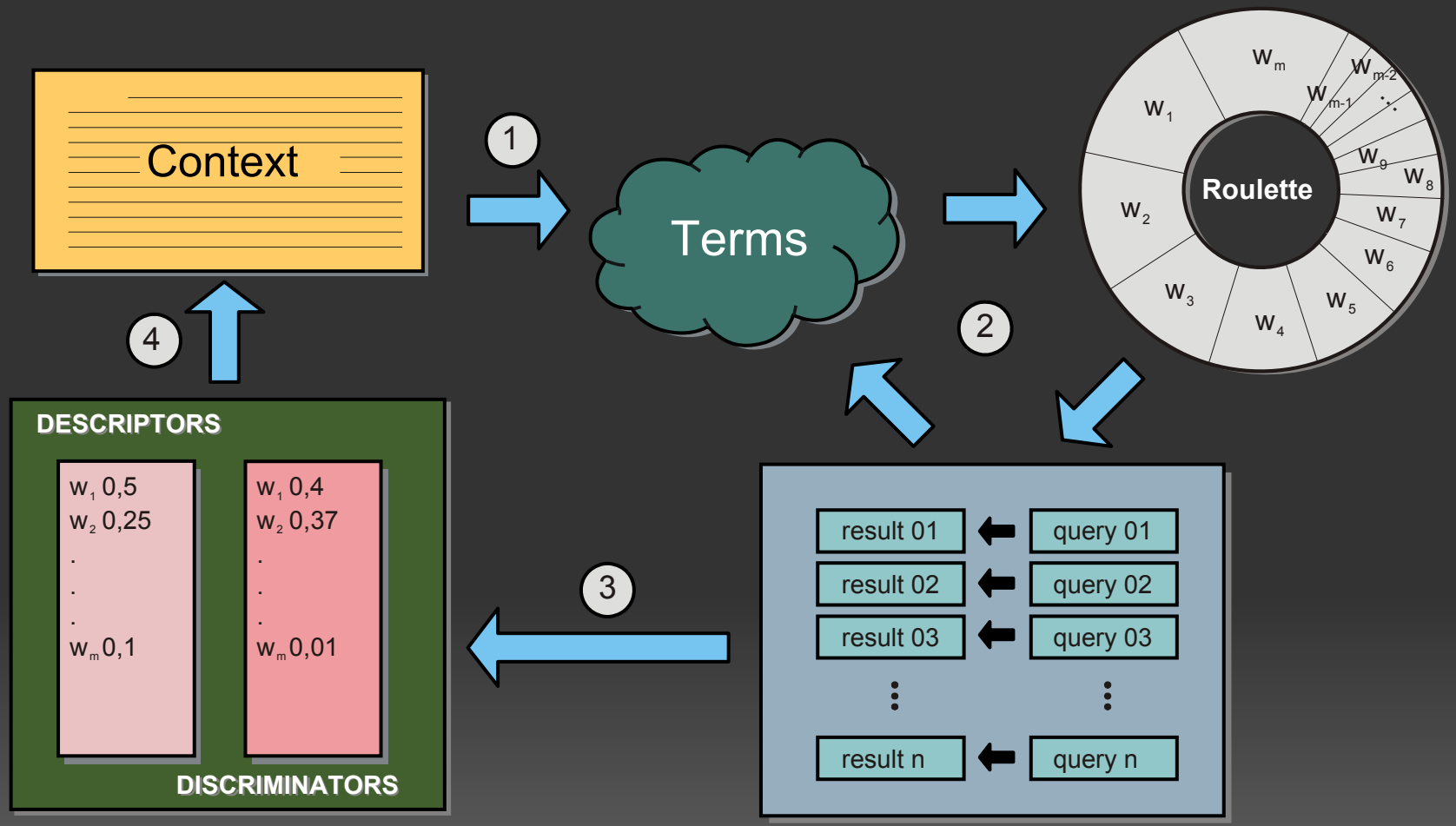
Initial Context		$\Delta(t_i, d_0)$
jvm	0	0,848
jdk	0	0,848
virtual	1	0,566
programming	3	0,566
machine	2	0,524
language	1	0,517
java	4	0,493
coffee	0	0,385
island	0	0,385
province	0	0,385

Topic: Java Virtual Machine

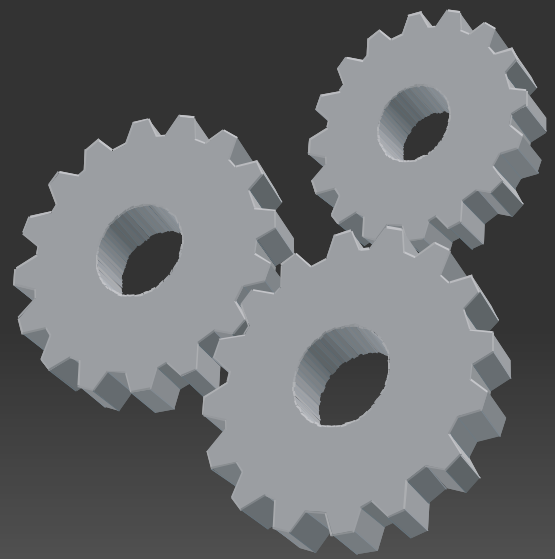
Term discriminating power in a topic of a document

$$\Delta(t_i, d_j) = \sum_{k=0,\ k \neq j}^{m-1} (\sigma(d_k, d_j) \cdot \delta(t_i, d_k)^2)$$

Proposed Algorithm



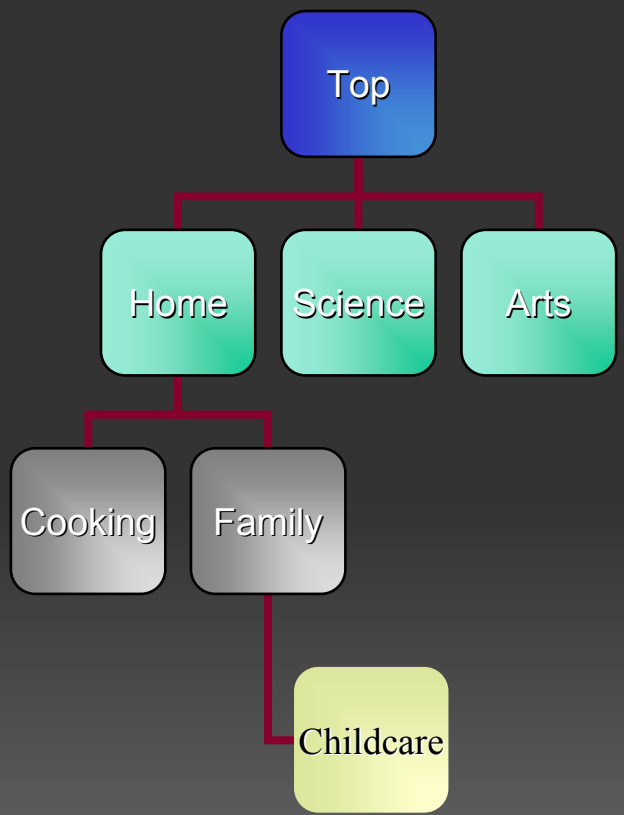
Evaluation



Evaluation

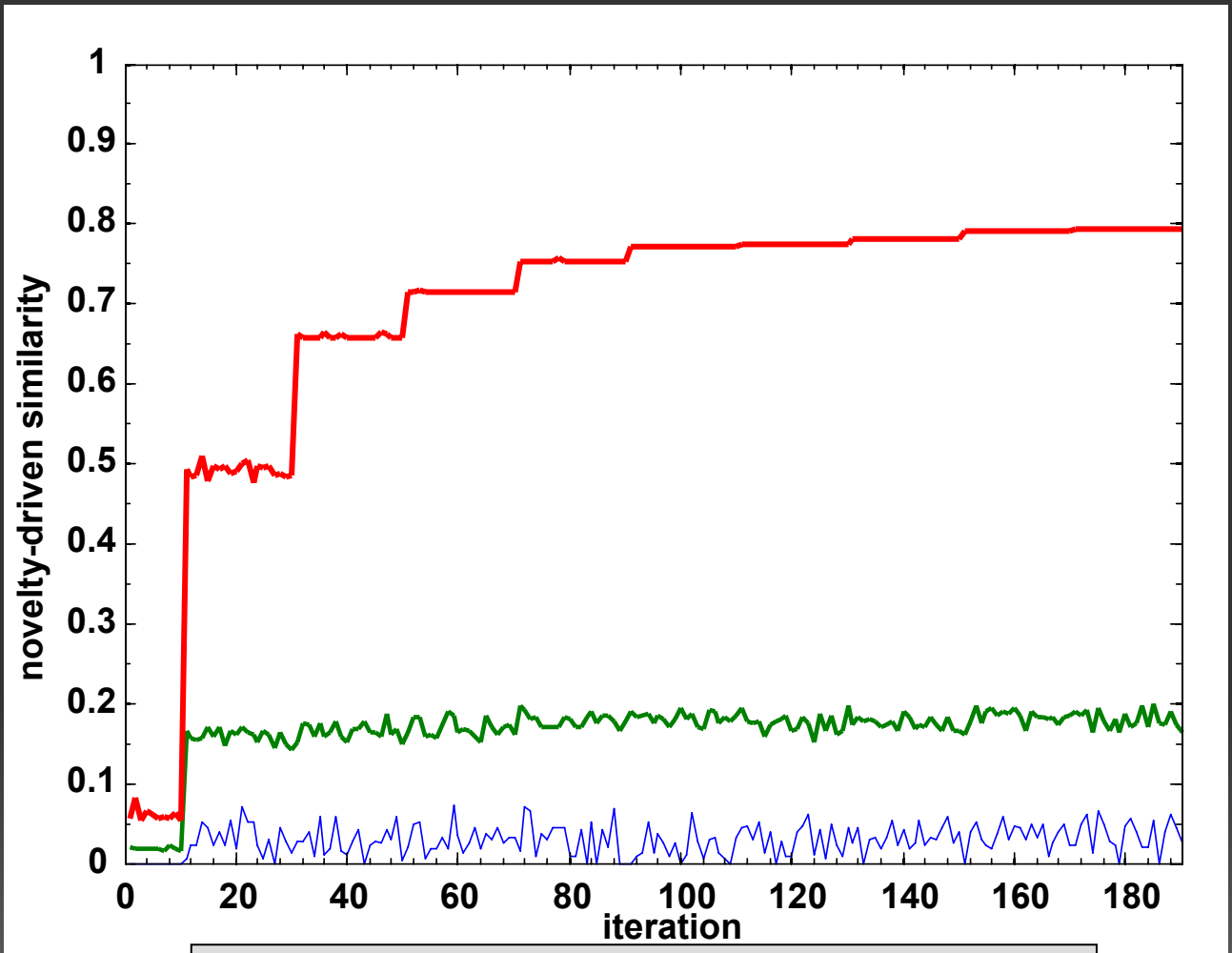
Local Index

- DMOZ – ODP Project
  - ~ 500 Topics
  - > 350,000 pages
  - English language
  - 3rd level of the hierarchy
  - 100 URLs at least



- Similarity
- Precision
- Recall

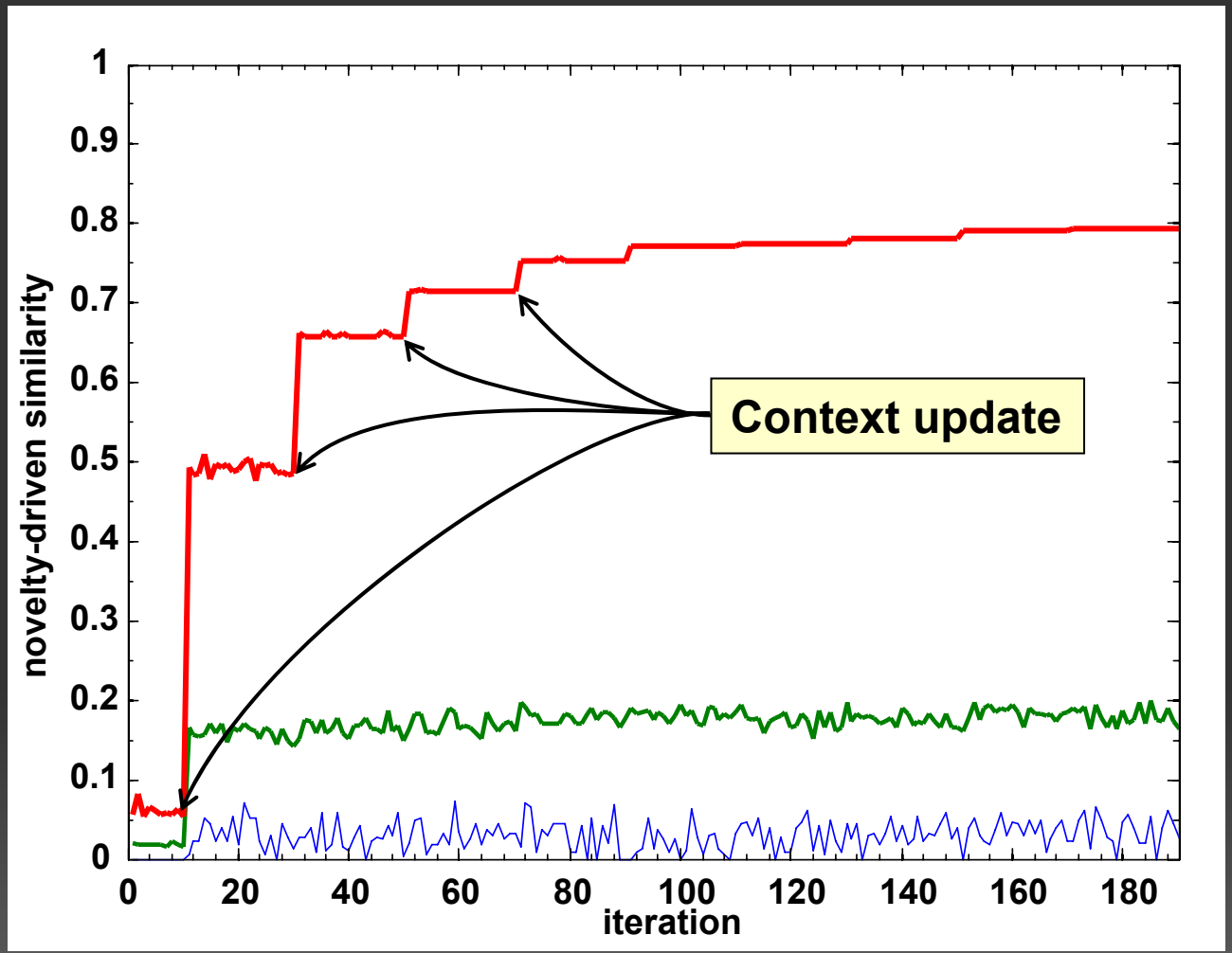
Evaluation – Similarity



Top/Computers/Open_Source/Software

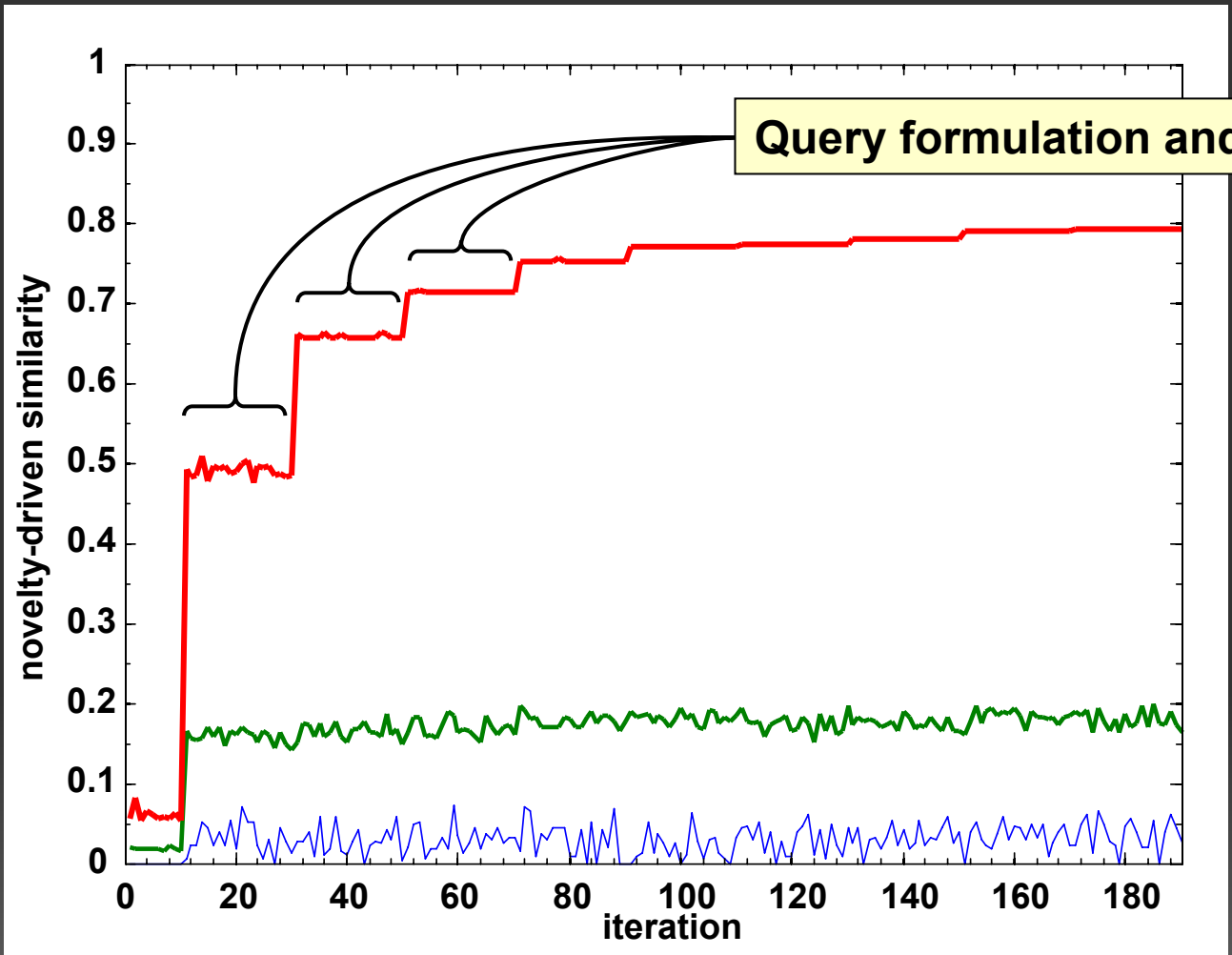
σ^N	Mean	95% CI
1st	0.0661	[0.0618; 0.0704]
best	0.5970	[0.5866; 0.6073]

Evaluation – Similarity



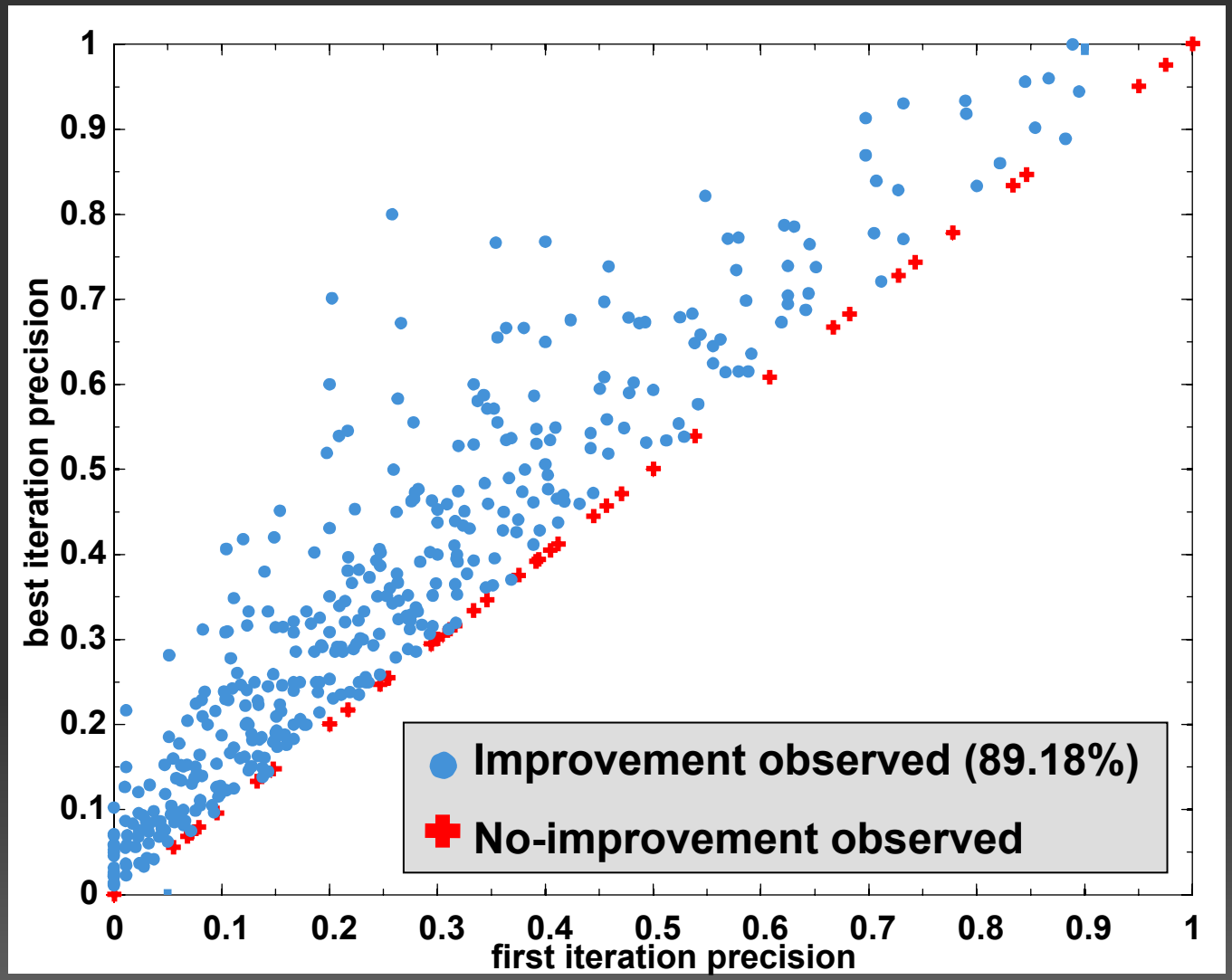
σ^N	Mean	95% CI
1st	0.0661	[0.0618; 0.0704]
best	0.5970	[0.5866; 0.6073]

Evaluation – Similarity

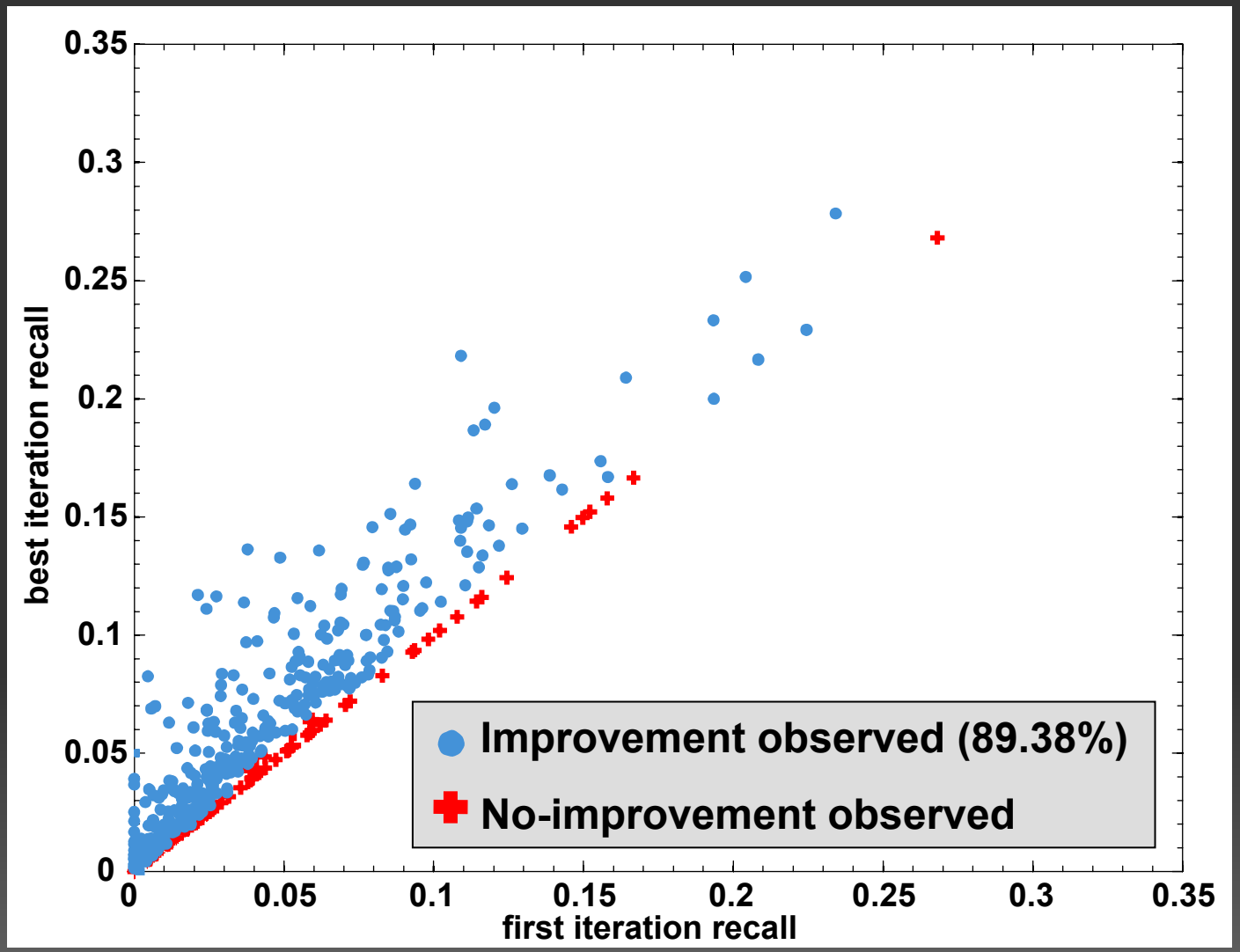


σ^N	Mean	95% CI
1st	0.0661	[0.0618; 0.0704]
best	0.5970	[0.5866; 0.6073]

Evaluation – Precision



Evaluation – Recall



Conclusions

- We presented an intelligent IR approach for learning *context-specific* terms.
  - Take advantage of the *user context.*
- We have shown evaluations and the *effectiveness* of incremental methods.

Future Work

- Investigate different parameters.
- Develop methods to learn and adjust parameters.
- Run additional tests to comparing with other methods.

Thank you!

Laboratorio de Investigación y Desarrollo en Inteligencia Artificial

lidia.cs.uns.edu.ar

Universidad Nacional del Sur
Bahía Blanca

www.uns.edu.ar

CONICET

AGENCIA