
Web Science – **Investigating the Future** of Information and Communication

Gianluca Demartini

with Claudiu S. Firan, Julien Gaugaz, Tereza Iofciu, Ralf Krestel, Wolfgang Nejdl, Arjen P. de Vries, Jakub Zakrzewski, and Jianhan Zhu

FINDING ENTITIES AND TRACING THEIR IDENTITY

Who I am

Gianluca Demartini

Intern working with Hugo

- LivingKnowledge project
- Entity / Novelty / over Time

M.Sc. from University of Udine, Italy (2005)

Ph.D. Student at L3S Research Center
University of Hannover, Germany (2006)

Research Interests:

- Entity Retrieval
- Semantic Web
- IR evaluation

Outline

Entity Retrieval: a Model and techniques

- In the Enterprise
- in Wikipedia

Entity Identity: Management over Time

(*Entity Retireval Evaluation: Stratified Pooling Techniques*)

ENTITY RETRIEVAL

Entity Ranking

Many users search for specific entities instead of just any type of documents

Ranking People

Expert Finding in TREC-ENT (Enterprise Track)

Collection:

- Corpus: crawl of *.w3.org sites
- People: names of 1092 people who may be experts

Query:

- ‘`information retrieval`’

Results:

- A **list of people** who know about information retrieval

Ranking Actors

Queries are lists of actors on the Web, e.g.

- Query: 1930s
  - Answers: Fred Astaire, Charlie Chaplin, W.C. Fields, Errol Flynn, Clark Gable, Greta Garbo, etc
- Query: action
  - Answers: Arnold Schwarzenegger, etc

Ranking...

People

- Expert Finding evaluation

Actors

- No evaluation initiative… yet?!

Car companies, countries, museums, …

[i.e., insert your fav entity type here]

Entity Ranking!!!

A Vector Space Model for Ranking Entities and Its Application to Expert Search (ECIR09)

Our contribution

A general model for ranking entities in a document collection

- Allowing integration of known techniques
- For any type of entity

An application to the expert finding task

The Model

Documents $D=d_1,...,d_m$

Entities $E=e_1,...,e_n$

Topics $T=t_1,...,t_l$

Query q

Rank $e_i \in E$ by degree of relevance to q

Documents as vectors in the VS

Documents as vectors in the VS

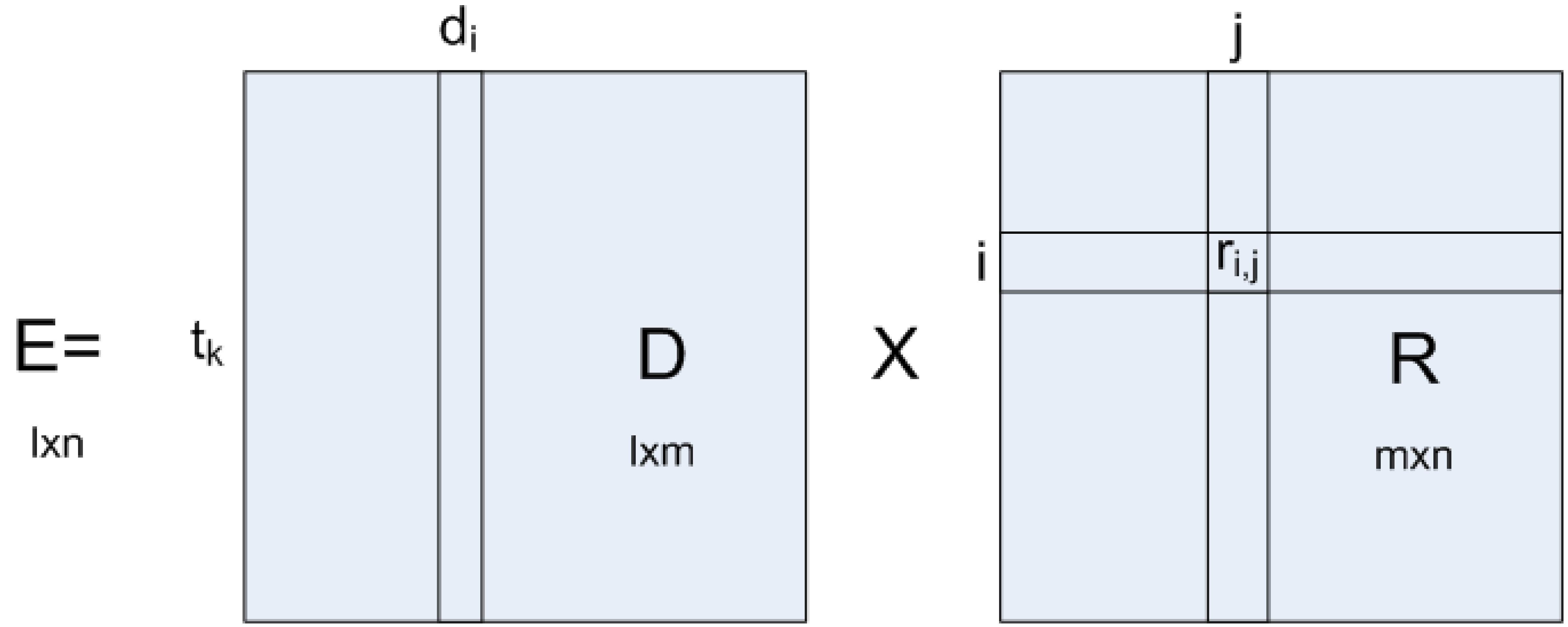
- $d_i = d_{1,i} t_1 + \ldots + d_{l,i} t_l$

Relationship between documents and entities

- $f : D \times E \rightarrow R : (d_i , e_j) \rightarrow r_{ij}$

Entities as vectors in the VS

- $$e_j = \sum_{k=1}^{l} \left(\sum_{i=1}^{m} d_{k,i} r_{i,j} \right) t_k$$



Query

Query $q=q_1t_1+...+q_nt_n$

Cosine similarity

$$sim(q,v)=\frac{q\cdot v}{\|q\|\|v\|}$$

- Where $v\in\{d_i,e_j\}$

Extensions

Document dependent

- $E = D \times (diag(x) \times R)$
- diag(x) is m x m with x_{ii} is the weight for d_i

Entity and Topic dependent

- $E' = E \circ W$
- W is l x n with w_{jk} is weight for e_j on t_k

Entity dependent

- $E'' = E' \times diag(cf)$
- diag(cf) is n x n and cf_{jj} is the cost of e_j

An application: Expert Search

We adapt the model to Expert Search task

- We fix the entity type to people
- The query describes desidered expertise

TRECent 2006

- W3C web sites
- 300k documents
- 1092 (official) candidate experts

Projection Similarity

Cosine sim does not favour long documents

We should favour experts with more expertise

$$projSim(q,v) = \cos\theta \|v\|$$

The longer the expert vector the higher sim

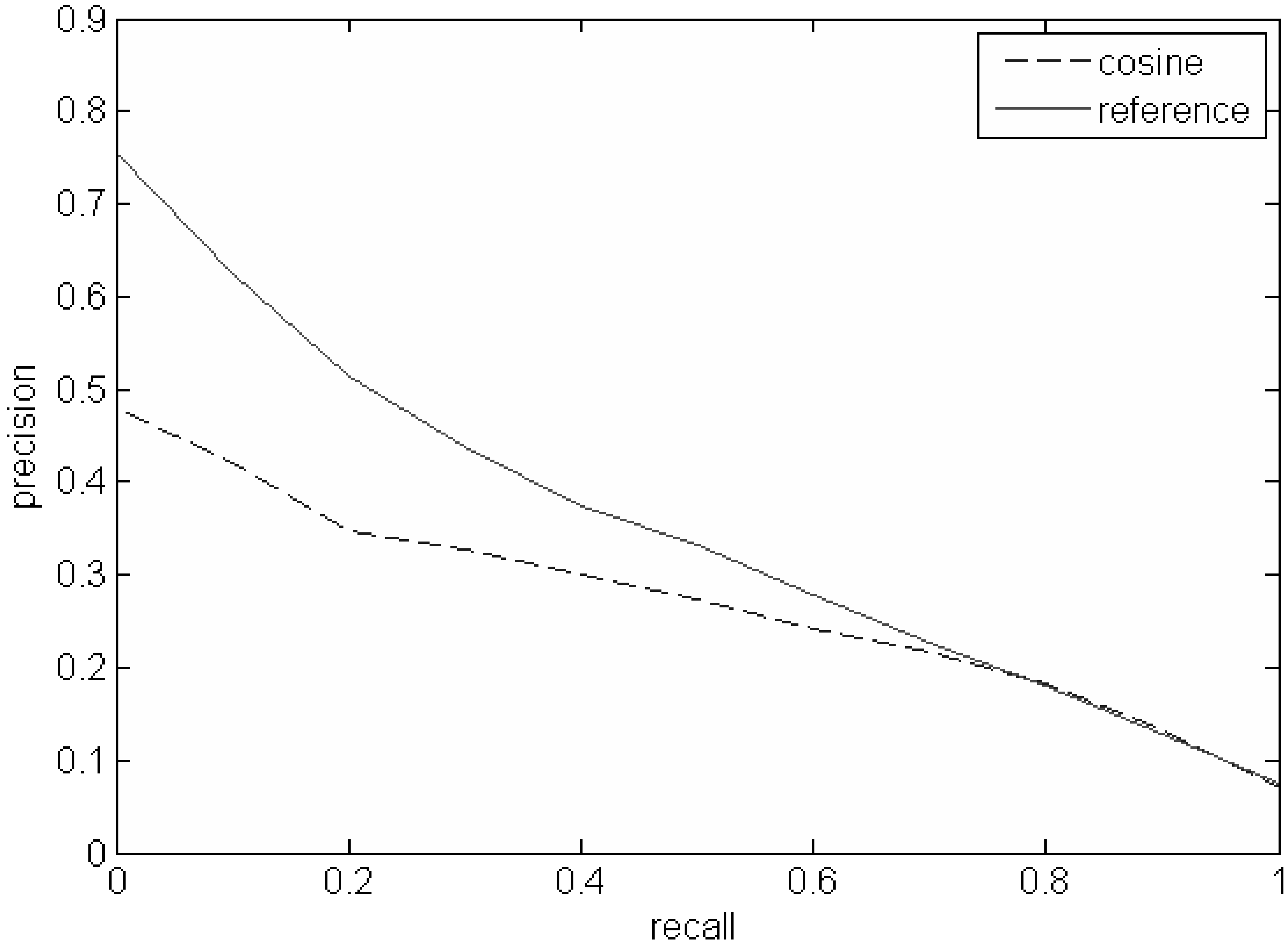
Experiments

Projection similarity for Expert Search

Explore

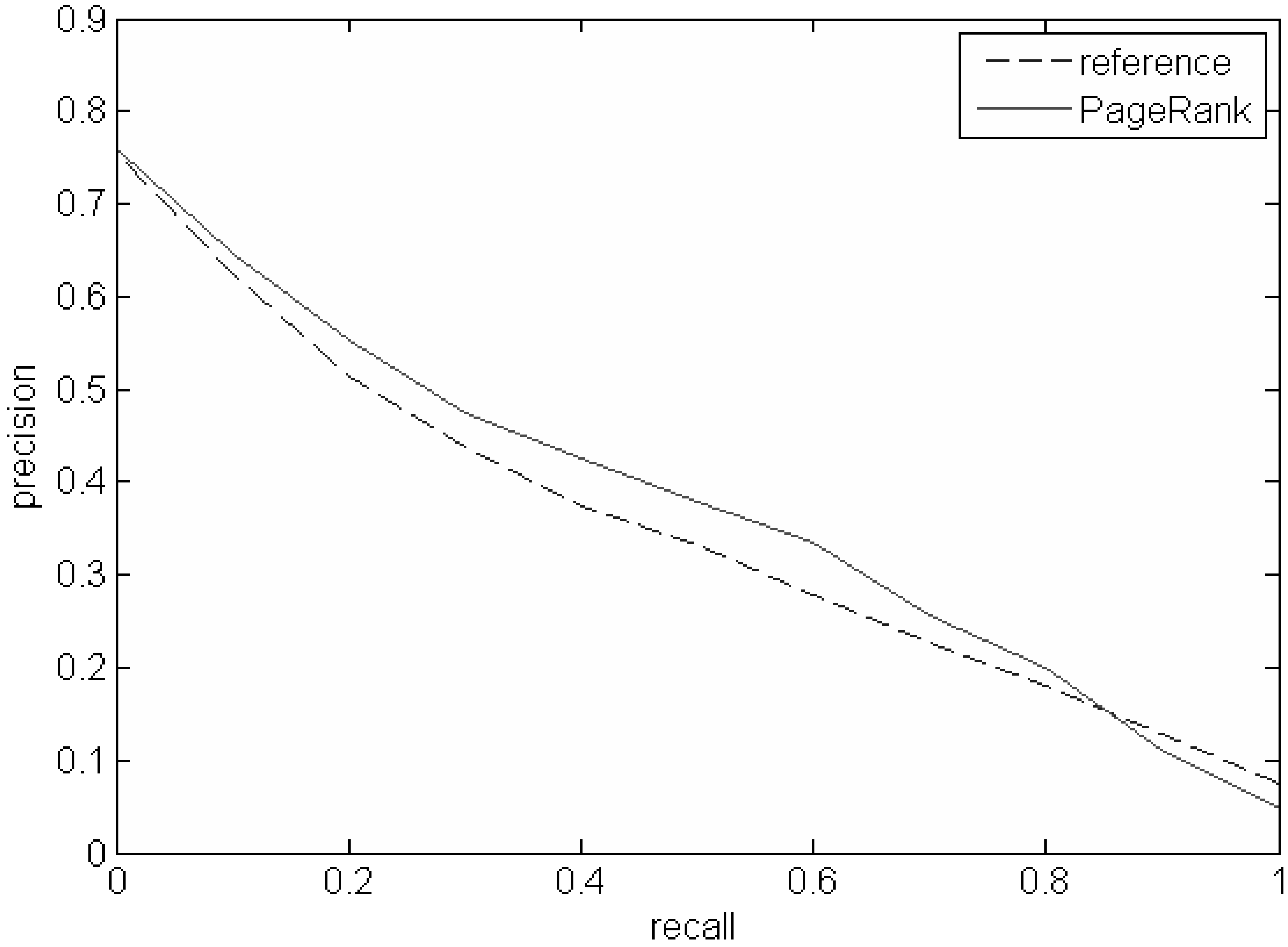
- document dependent extensions
- different space dimensions

ProjSim vs CosineSim



$$E = D \times (diag(x) \times R)$$

Document dependent extension



Vector Space Dimensions

Dimension	Term	LSA	LexComp	LexComp Pruned
MAP (p-value)	0.3370	0.0894 (p =0.0)	0.3586 (p =0.5927)	0.3625 (p =0.5374)

$$\{\ adjective?\ noun+\ \}$$

Related Work

Expert Finding

- Balog's model 1 [Balog et al. SIGIR06]
- Voting Model [Macdonald and Ounis CIKM06, ECIR07, ECIR08]
- Experise evidence [Macdonald et al. ECIR08]
- Topic drift: ProjSim allows multiple expertises

Conclusions

We presented a model for Entity Ranking

- It is based on the VSM
- Can be applied where entities are available
- Can be extended with different types of evidence

We applied to the task of Expert Finding

- By use of a custom similarity measure
- Exploring different extensions

Next steps:

- Perform the Entity Ranking task in a web collection

APPROACHES TO ENTITY RETRIEVAL IN WIKIPEDIA

Possible approaches to XER

Link structure [Pehcevski et al. ECIR08]

Language Models [Weerkamp et al. INEX08]

Passage retrieval [Zaragoza et al. CIKM07]

It is a recent task (2y): low effectiveness

All previous work use category information

INEX Wikipedia Collection

INitiative for the Evaluation of XML Retrieval

English Wikipedia 2006

659,338 articles

XML version preserving structural and typographical tags

25+35 topics (queries) created and assessed by the participants

Topic 60

Q

Title
olympic classes dinghy sailing

Xs

Entities
470 (dinghy) (#816578)
49er (dinghy) (#1006535)
Europe (dinghy) (#855087)

T_X

Categories
dinghies (#30308)

Description
The user wants the dinghy classes that are or have been olympic classes, such as Europe and 470.

Narrative
The expected answers are the olympic dinghy classes, both historic and current. Examples include Europe and 470.

Algorithms

- Structure based techniques (WISE08)
  - Using outgoing links
  - Lexical compounds
- NLP based techniques (LA-WEB08)
  - Synonyms and Related Words
  - Query extension: synonyms of nouns in the Keywords + Word Sense Disambiguation for the correct meaning

Baseline Query

Page text ⇔ Topic title

Page categories ⇔ Topic categories

Outgoing Links

Outgoing links of Wiki pages = concise information about the key concepts

- Nicolas Bloembergen:
  - Dutch
  - Physicist
  - American
  - Harvard University
  - 1948
  - …
  - University of Utrecht
  - Nuclear magnetic resonance
  - Lorentz Medal
  - Nobel Prize in Physics
  - Laser spectrology

Search in these “outgoing links” additionally to the full text

Lexical Compounds

- Find expressions of the form:

{ adjective? noun+ }

[Hybrid cars] ~~sold in~~ [Europe]

- Search with them instead of the full text

Entity Ranking Algorithms

- Synonyms
- Related Words (other than syn.)
- Core Characteristics
  - Clean the Keywords removing terms (and synonyms) appearing in Category
  - Keep only nouns and adjectives in Keywords
- Named Entities
  - Use only NE (i.e., organizations, locations, persons) from Keywords

Title	Tom Hanks movies where he plays a leading role.
Category	Films
Synonyms	Tom ”Uncle Tom” Hanks ”Thomas J. Hanks” movies film flick ”motion picture” ”motion-picture show” ”moving picture” pic picture ”picture show” ”moving-picture show” where he plays a leading role
Related Words	**Synonyms** plus 50 additional concepts related mainly to motion pictures
Core Characteristics	Tom Hanks leading role
Named Entities	Tom Hanks

Nr	Query; $q = \{category, W^C\} \cup \ldots$	xInfAP	P@10
1	$\{text, W^T\}$	0.2350	0.3057
9	$\{text, W^T\}, \{outLinks, W^T\}$	0.2556*	0.3371*
10	$\{text, W^T\}, \{outLinks, CC(W^T)\}$	0.2511	0.3114
11	$\{text, W^T\}, \{outLinks, NE(W^T)\}$	0.2504*	0.3171
12	$\{LC(W^T)\}$	0.2284	0.2971
13	$\{text, W^T \cup LC(W^T)\}$	0.2506	0.3257
14	$\{text, W^T \cup LC(W^T)\}, \{outLinks, W^T \cup LC(W^T)\}$	0.2616	0.3457
15	$\{text, W^T \cup SY(W^T)\}$	0.2439*	0.3257
16	$\{text, W^T \cup RW(W^T)\}$	0.2398	0.3199
17	$\{text, W^T \cup CC(W^T)\}$	0.2509*	0.3257
18	$\{text, W^T \cup NE(W^T)\}$	0.2530*	0.3257
19	$\{text, W^T \cup SY(W^T) \cup RW(W^T) \cup CC(W^T) \cup NE(W^T)\}$	0.2705*	0.3571*
20	$\{text, W^T \cup SY(W^T) \cup RW(W^T) \cup CC(W^T) \cup NE(W^T)\}$, $\{outLinks, CC(W^T)\}$	0.2682*	0.3599*
21	$\{text, W^T \cup SY(W^T) \cup RW(W^T) \cup CC(W^T) \cup NE(W^T)\}, \{category, W^T\}$	**0.2909***	**0.3971***
22	$\{text, +W^T \cup SY(W^T) \cup RW(W^T) \cup CC(W^T) \cup NE(W^T)\}$	0.0813*	0.1124*
23	$\{text, W^T \cup SY(W^T) \cup RW(W^T) \cup +CC(W^T) \cup NE(W^T)\}$	0.2627	0.3857
24	$\{text, W^T \cup SY(W^T) \cup RW(W^T) \cup CC(W^T) \cup NE(W^T)\}$, $\{outLinks, CC(W^T)\}, \{title, -W^T\}$	0.2748*	0.3657*
25	$\{text, W^T \cup SY(W^T) \cup RW(W^T) \cup CC(W^T) \cup NE(W^T)\}$, $\{outLinks, CC(W^T)\}, \{title, -W^C\}$	0.2534	0.3314

Conclusions

Entity Ranking must be tackled differently than traditional Information Retrieval

The use of simple Natural Language Processing & Link Analysis improves retrieval

Overall improvement in AP of 24%

Demo

- Spanish dishes
- http://okkam.l3s.uni-hannover.de:8080/er08web/
- http://search.yahoo.com
- http://www.google.com/squared
- http://correlator.sandbox.yahoo.com
- http://www.powerset.com

HOW TO TRACE ENTITY IDENTITY (ESWC 2009)

Entity Identity on the Web

Entity Name System (ENS)

- Provides globally uniques URIs given an entity description

Identity evolves over time

- One entity can have more than one identifier
  - http://dbpedia.org/resource/Tim_Berners-Lee
  - http://data.semanticweb.org/person/tim-berners-lee
- One identifier can refer to more than one entity
  - http://dbpedia.org/page/Paris

Operations on entity identifiers

Identity Decision Revision (IDR):

- Creation: new entity discovered
- Split: the representation describes two real world entities
- Merge: two descriptions about the same real world entity

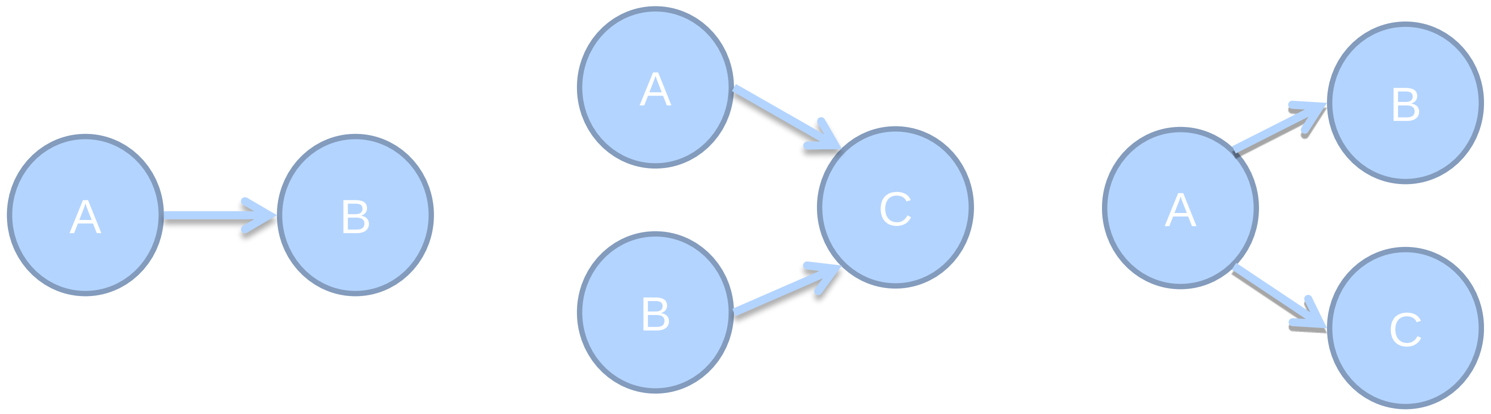
In Wikipedia:

- Page creation
- Redirection page: merge
- Disambiguation page: split

We propose to label URIs with lineage information

- Lineage Preserving ID (LPID)

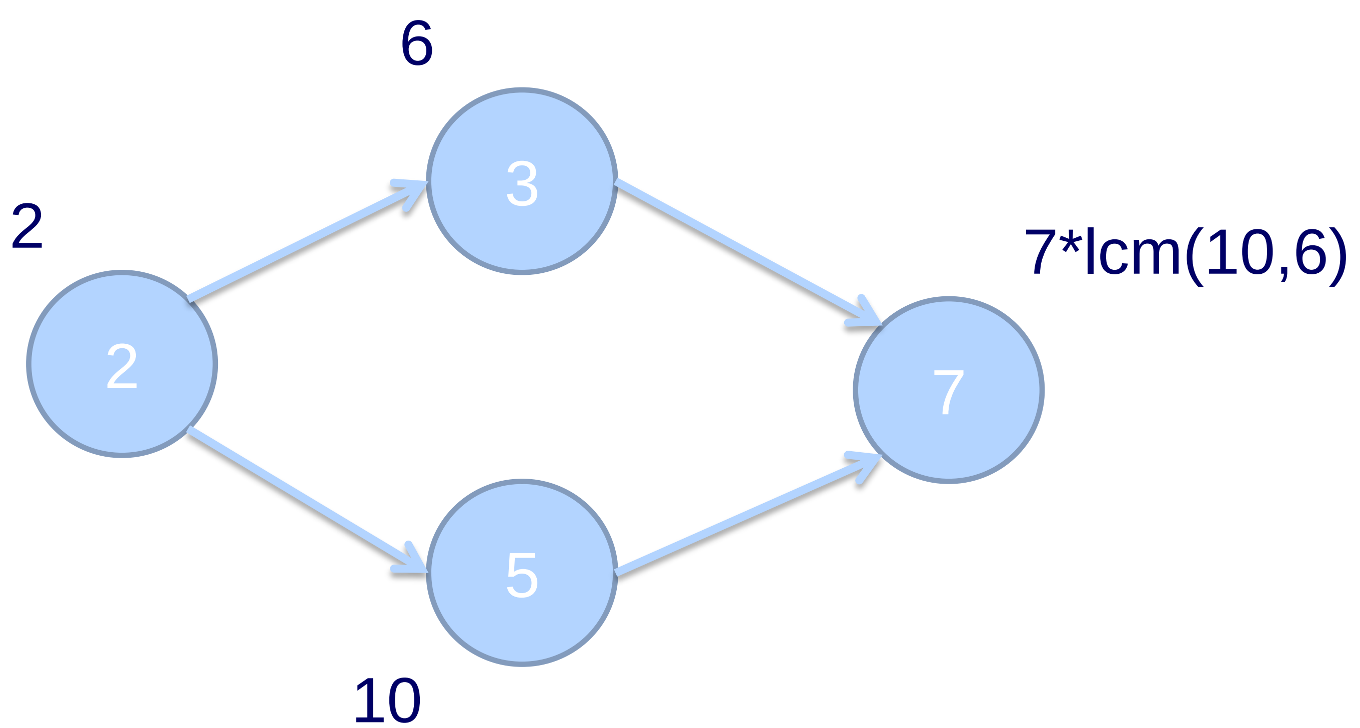
Identity revision graph



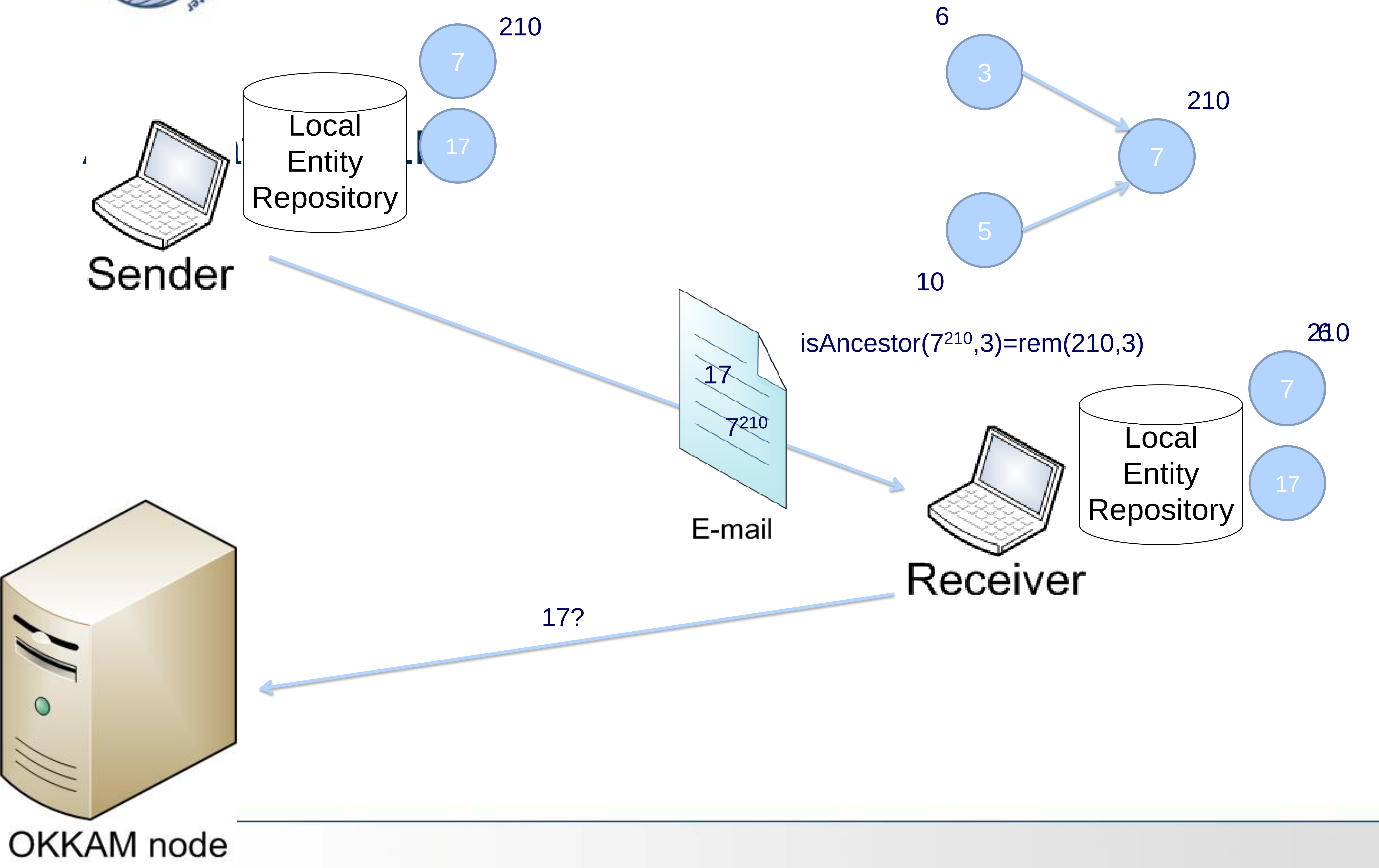
Deprecation Merge Split

Prime Numbers Labelling Scheme for DAGs

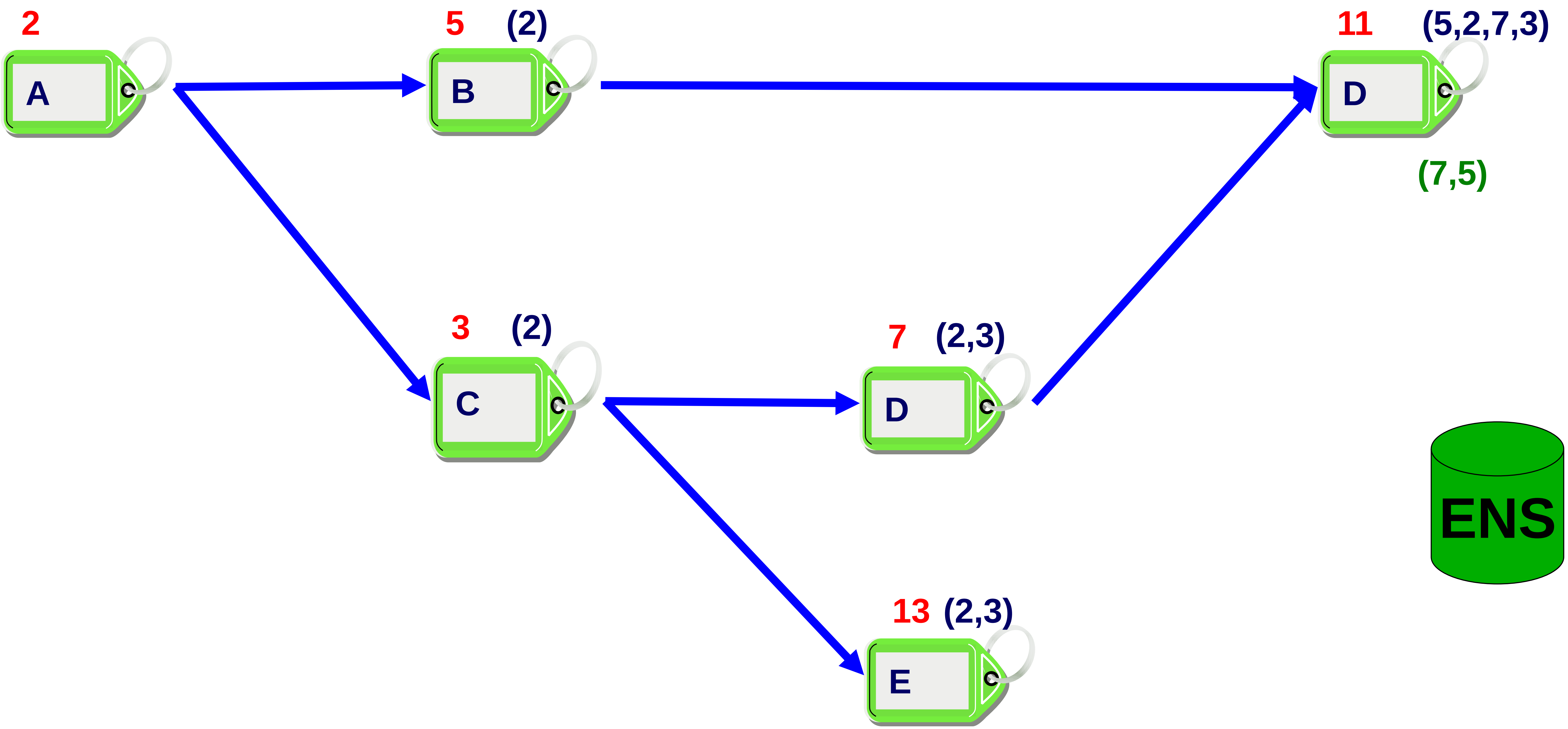
- DAG: G(V,E)
- Algorithm:
  - Assign a unique prime number *p* to each *v* in *V* **self-label**
  - Label each *v* with (*p** the least common multiplier of its ancestors' label) **ancestor-label**



Adapting Prime Number Labeling Scheme for Directed Acyclic Graphs
G Wu, K Zhang, C Liu, J Li - Database Systems for Advanced Applications, 2006 Springer


210
7
17
Local Entity Repository
Sender
6
3
210
7
5
10
isAncestor(7210,3)=rem(210,3)
17
7210
E-mail
210
7
17
Local Entity Repository
Receiver
17?
OKKAM node

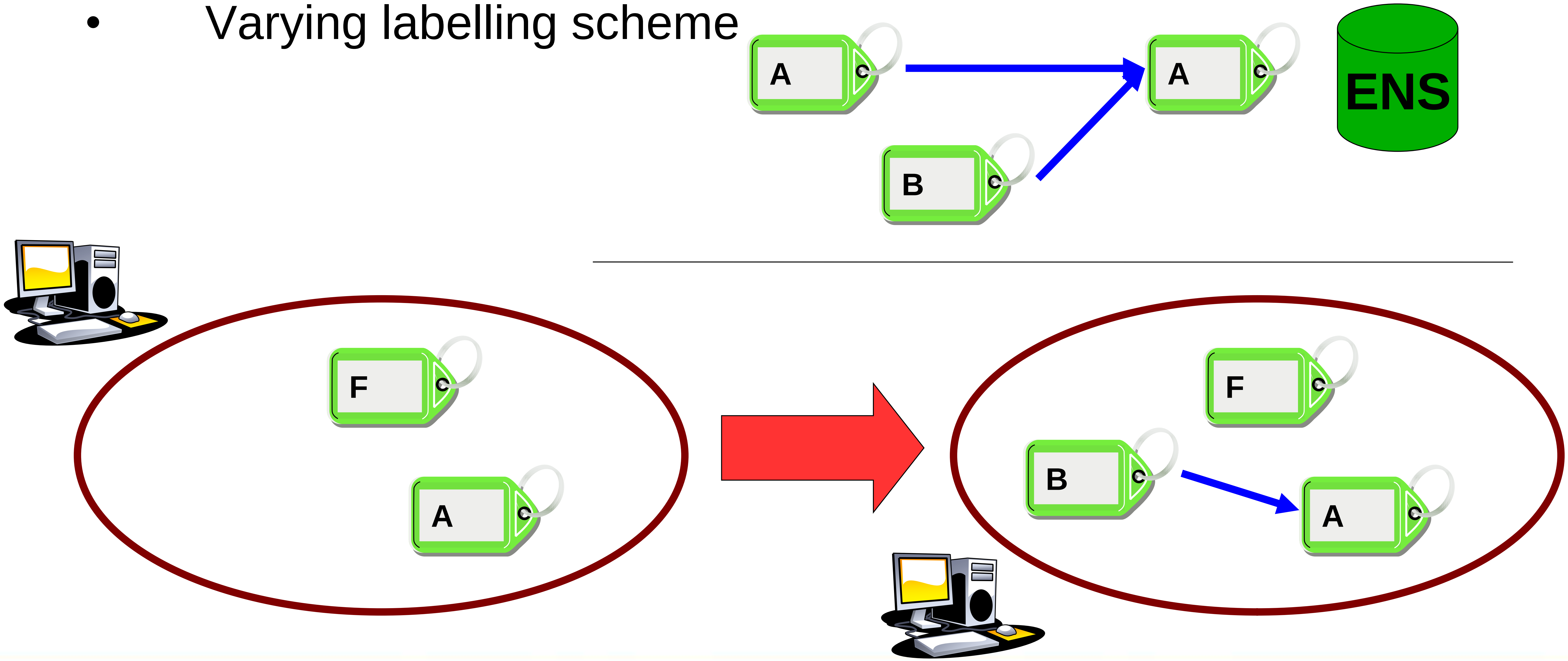
Labelling URIs with List of Ancestors



List labelling

Experiment Scenario

- A client is receiving descriptions about different entities
- Based on the URI of the entities and on their lineage annotation, it decides what description to ask an update for
- Varying labelling scheme



Overview of Labelling Schemes

- Baseline: no labelling
  - Always update all URIs
- Prime Number Labelling
  - Update only URIs which are ancestors in the local repository
- List labelling
  - Equivalent to prime number labelling
  - Stores the ancestors as list instead of as lcm(ancestors)

What to measure

Identifier quality

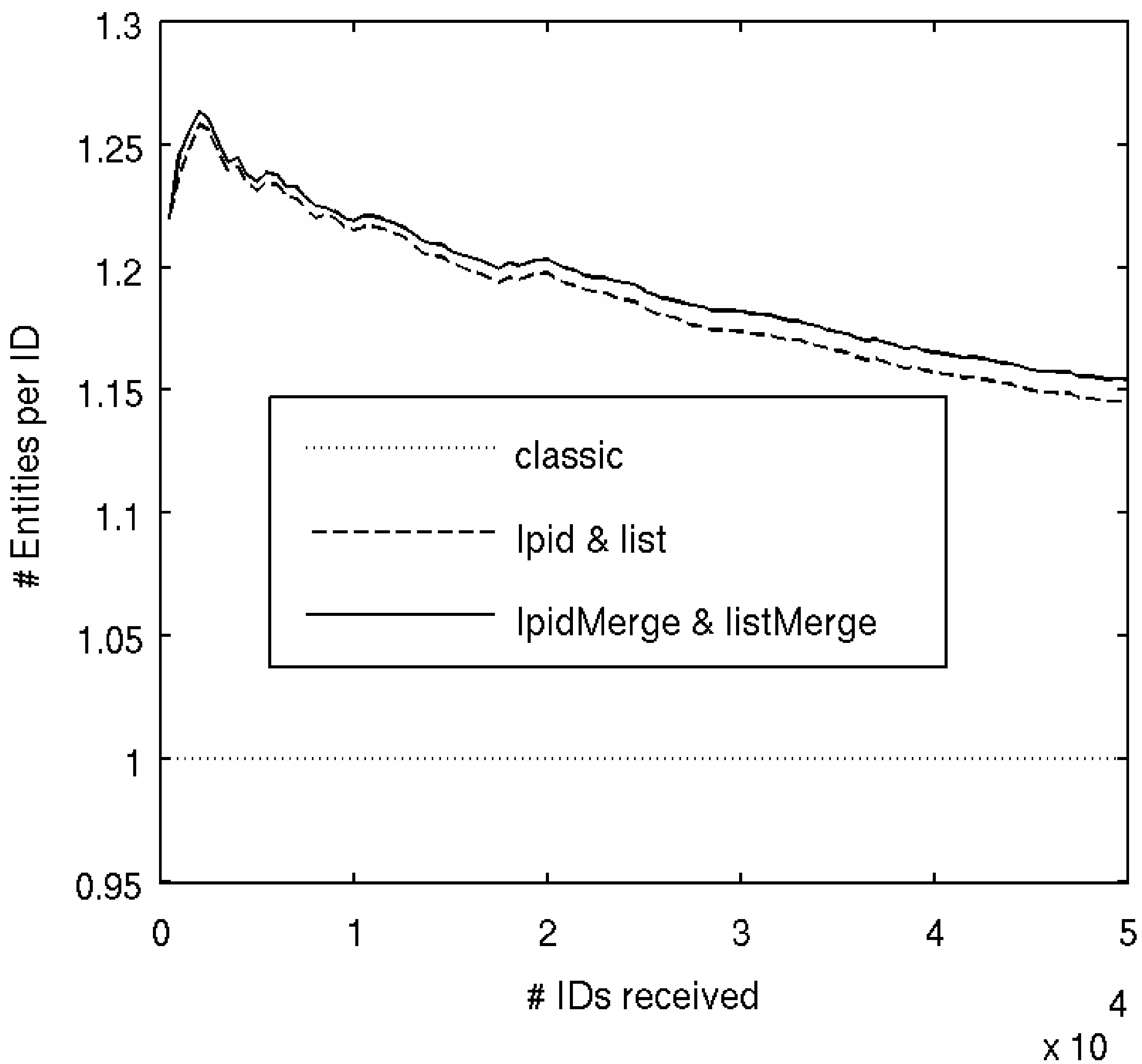
- entities per ID

Network Traffic

- size of metadata

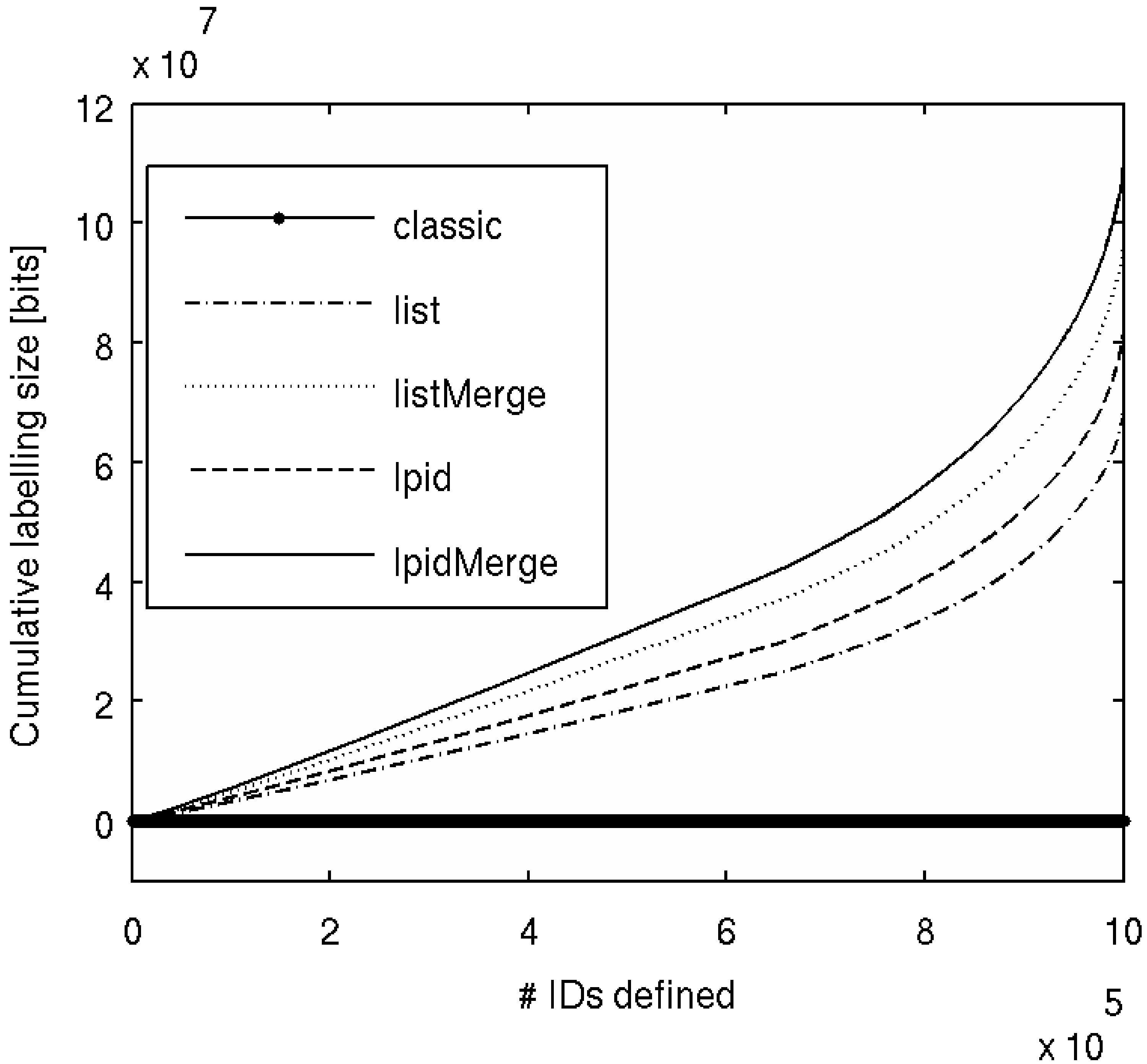
Time for update

Identity Quality



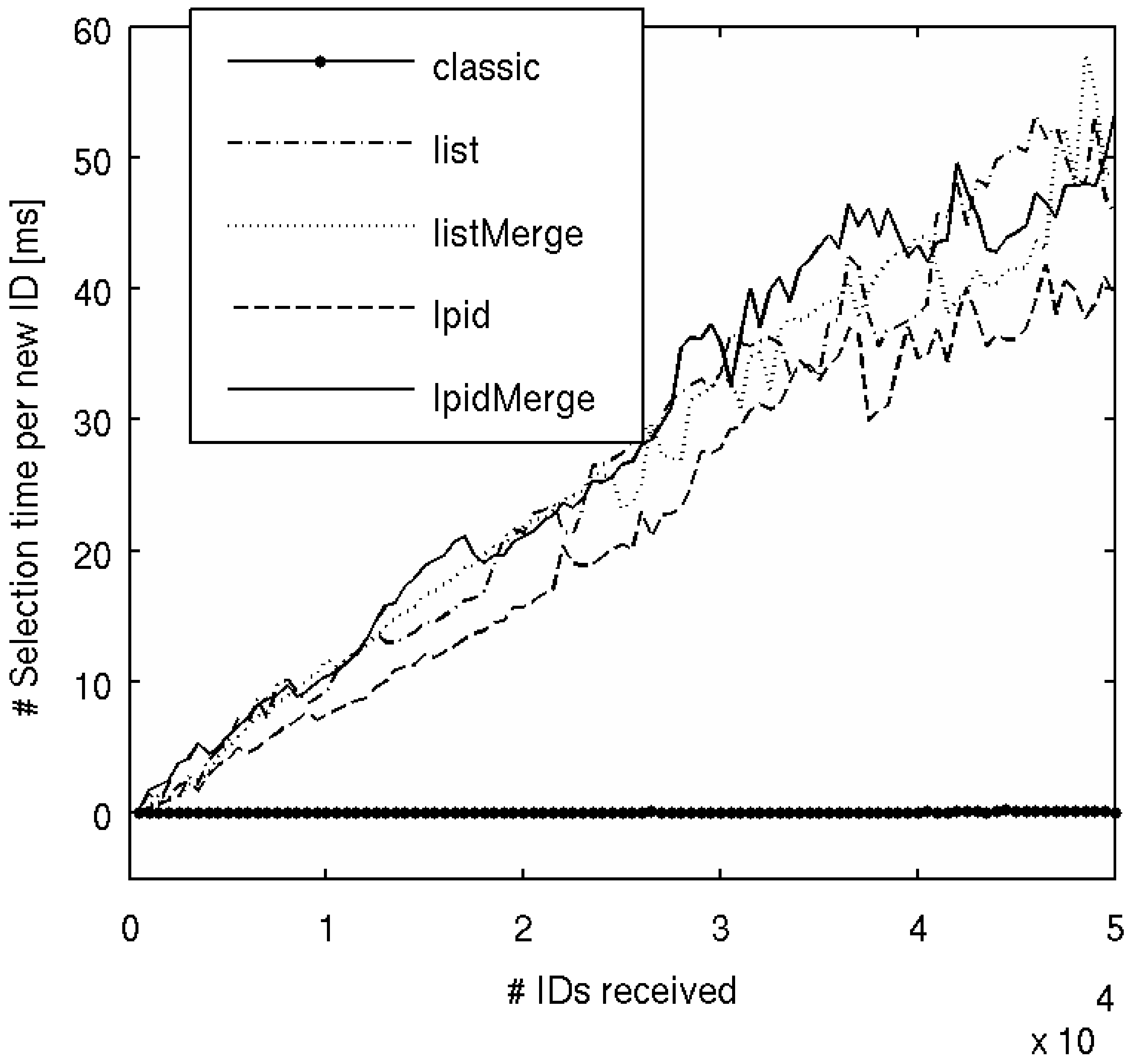
Average number of Entities in the ENS identified by a URI in the client
On Artificial Evolution Graph

Labelling Size

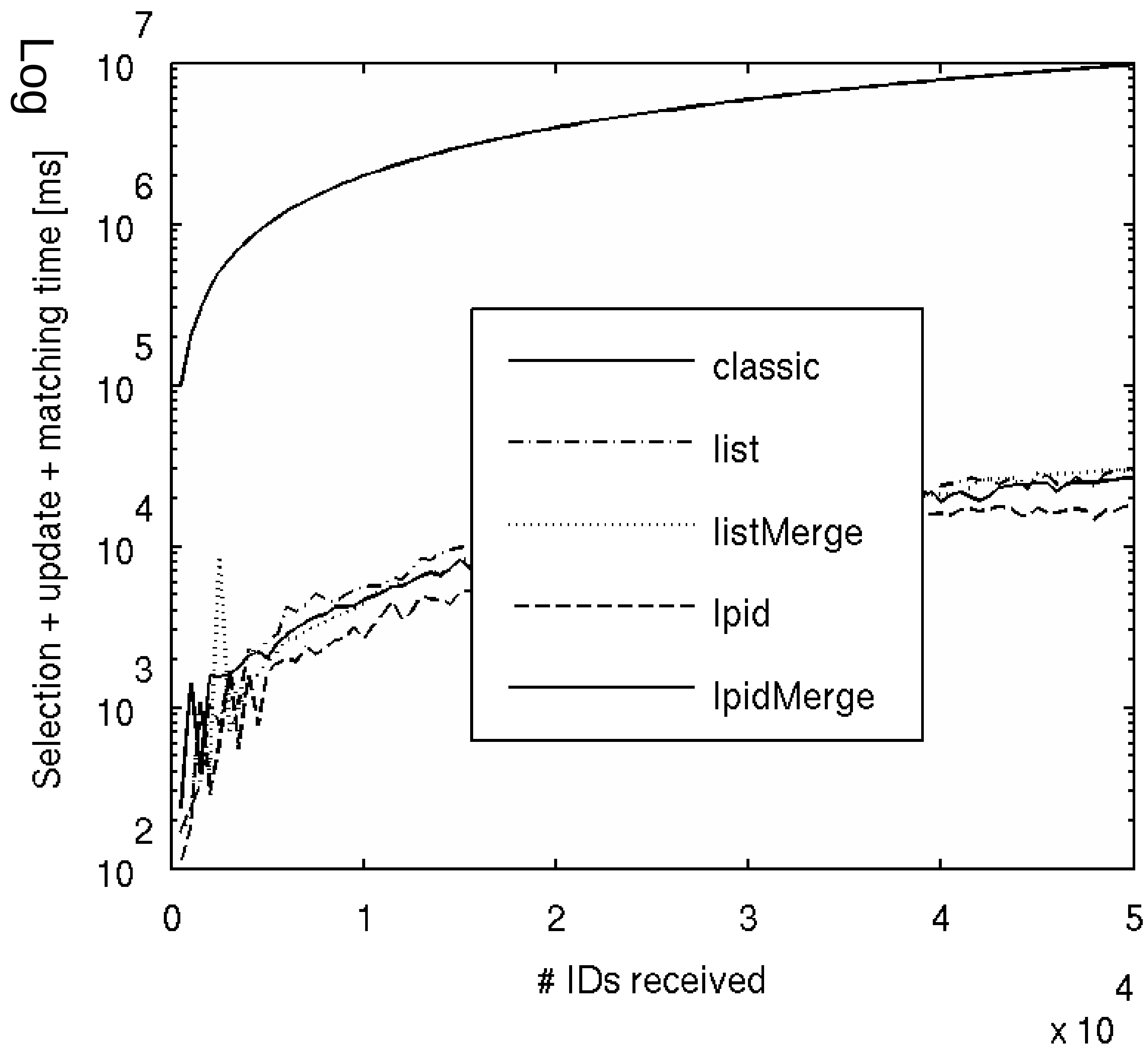


13 MB

Performances



Selection time on the Artificial Wikipedia



Total time (= selection + transmission + matching)
for updating the client's repository
On Dutch Wikipedia

Experiment Results Summary

- URI lineage labelling dramatically reduces the number of update requests to the ENS server
- URI lineage labelling provides a lower identity quality, but still a reasonable one

Conclusion

- Using id lineage labelling does allow to considerably reduce network traffic and server workload while keeping the identifier's quality reasonable.

INEX XER – Entity Retrieval Evaluation

XML Entity Ranking

Topical query Q

Entity (result) type T_X

A list of entity instances Xs

Systems employ XML element text, structure, links

Not relevant for XER…

Articles *on topic* are not necessarily relevant entities

- Actually, they are surprisingly often not!

- INEX 2007 adhoc-derived XER topics show that only about 35% out of original relevant documents have been assessed as relevant

Topic 60

Q

Title
olympic classes dinghy sailing

Xs

Entities
470 (dinghy) (#816578)
49er (dinghy) (#1006535)
Europe (dinghy) (#855087)

T_X

Categories
dinghies (#30308)

Description
The user wants the dinghy classes that are or have been olympic classes, such as Europe and 470.

Narrative
The expected answers are the olympic dinghy classes, both historic and current. Examples include Europe and 470.

Example 2008 Topics

Countries that have hosted FIFA Football World Cup tournaments: *countries; football world cup*

Formula 1 drivers that won the Monaco Grand Prix: *racecar drivers; formula one drivers*

Italian nobel prize winners: *nobel laureates*

…

Many examples on
http://www.ins.cwi.nl/projects/inex-xer/topics/

Google Sets

Topic 60

Title

olympic classes dinghy sailing

Entities

470 (dinghy) (#816578)

49er (dinghy) (#1006535)

Europe (dinghy) (#855087)

Categories

dinghies (#30308)

Description

The user wants the dinghy classes that are or have been olympic classes, such as Europe and 470.

Narrative

The expected answers are the olympic dinghy classes, both historic and current. Examples include Europe and 470.

Predicted Items
49er
470
europe
laser
~~optimist~~
finn
~~420~~
tornado
yngling
star
laser radial
~~29er~~
~~snipe~~
~~mistral~~
~~contender~~

INEX XER Overview 2008

2008 Tasks

Entity Ranking (ER)

- Given Q and T, provide Xs

List Completion (LC)

- Given Q and Xs[1..m]
- Return Xs[m+1..N]

INEX XER 2008 Assumptions

Entities (Xs) are represented as Wikipedia pages

Binary relevance, MAP (`xinfAP*`)

* A simple and efficient sampling method for estimating AP and NDCG. Emine Yilmaz, Evangelos Kanoulas, and Javed A. Aslam. SIGIR'08

Runs

Participation in 2008

- >60 groups sign up
- 11 groups submit topics
- 6 groups submit 33 runs

- 12 groups assess topics

Pooling by Sampling

Approaches:

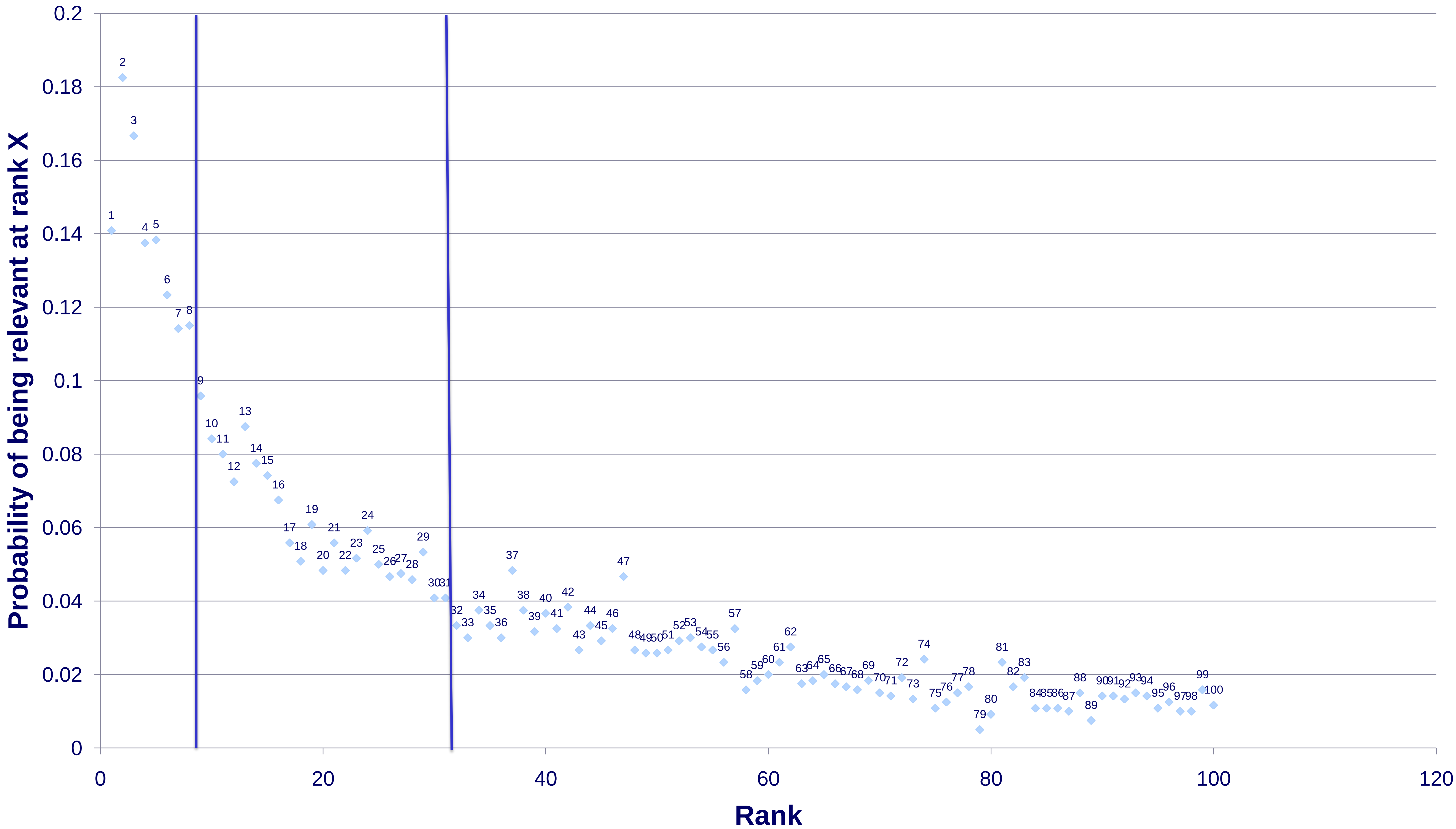
- Random sampling
- Relevance based sampling
- Stratified sampling

Collection

- 24 XER2007 topics (pool size: 50)

Comparison

- IRSs ranking changes with less assessments

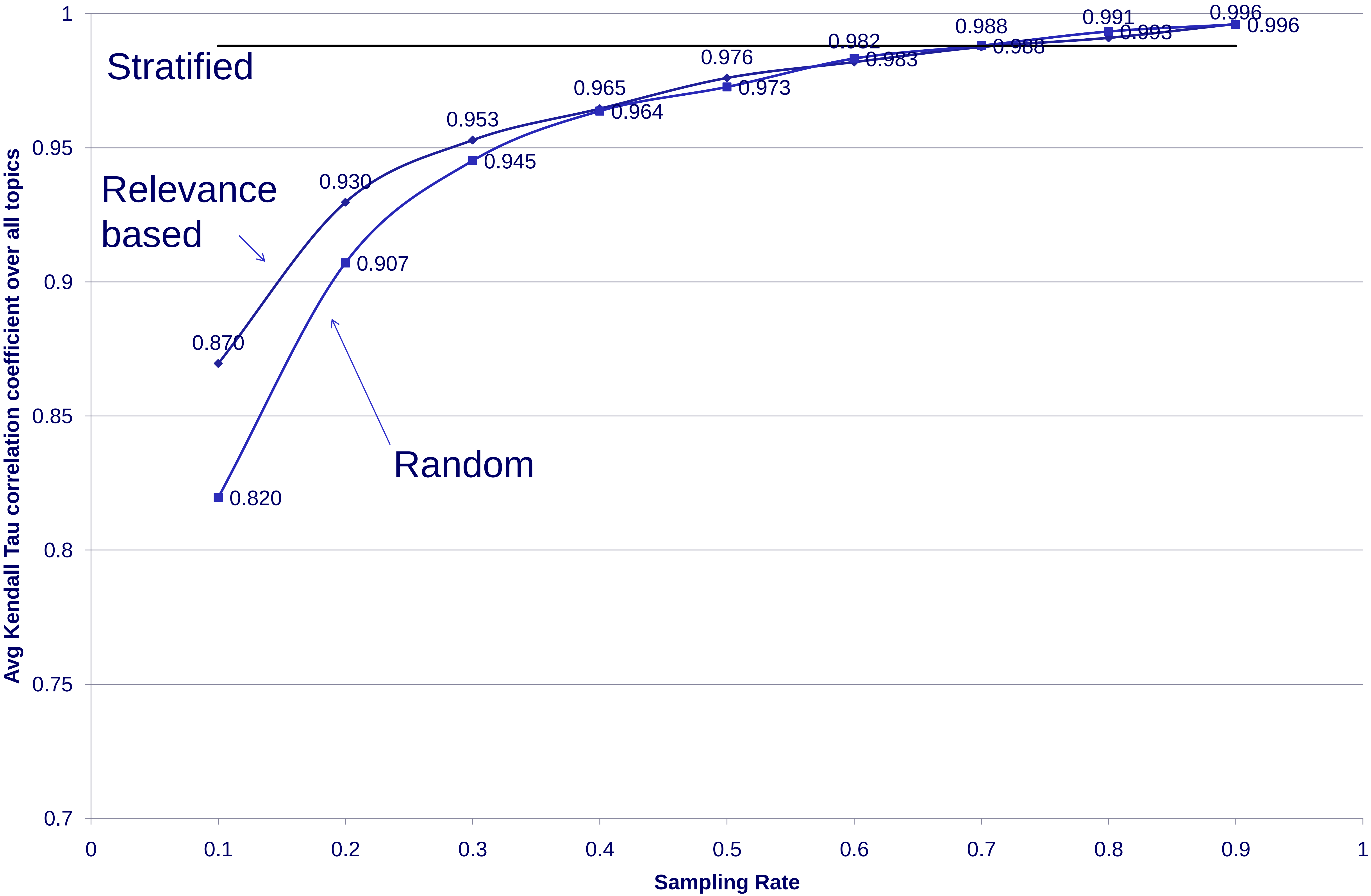

Probability of being relevant at rank X
0.2
0.18
0.16
0.14
0.12
0.1
0.08
0.06
0.04
0.02
0
0
20
40
60
80
100
120
Rank

Stratified Sampling

{ 1,8 } 100%

{ 9,31 } 70%

{ 32,100 } 30%



Stratified
Relevance based
Random
Avg Kendall Tau correlation coefficient over all topics
Sampling Rate
0.870
0.930
0.953
0.965
0.976
0.982
0.988
0.991
0.996
0.820
0.907
0.945
0.964
0.973
0.983
0.988
0.993
0.996
0.7
0.75
0.8
0.85
0.9
0.95
1
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1

Pool Contribution

Random/Relevance based Sampling:

- at 70%: 35 docs out of top 50

Stratified Sampling:

- 45 docs out of top 100 (30 docs out of top 50)

XER 2007 pool: 50 docs

INEX 2009

New Annotated Wikipedia collection

1. register at the INEX website
   http://www.inex.otago.ac.nz/people/register.asp
   demartini@L3S.de

2. index the provided Wikipedia collection

3. design an algorithm for finding entities

4. run the set of queries and produce your runs

Timeline:

- - 04 Oct - confirm your participation
- - 15 Nov - run submission + textual descriptions of runs
- - 23 Nov - INEX pre-proceeding papers due

References

Gianluca Demartini, Julien Gaugaz, Wolfgang Nejdl: A Vector Space Model for Ranking Entities and Its Application to Expert Search. ECIR 2009: 189-201

Gianluca Demartini, Claudiu S. Firan, Tereza Iofciu, Wolfgang Nejdl: Semantically Enhanced Entity Ranking. WISE 2008: 176-188

Gianluca Demartini, Claudiu S. Firan, Tereza Iofciu, Ralf Krestel, Wolfgang Nejdl: A Model for Ranking Entities and Its Application to Wikipedia. LA-WEB 2008: 29-38

Julien Gaugaz, Jakub Zakrzewski, Gianluca Demartini, Wolfgang Nejdl: How to Trace and Revise Identities. ESWC 2009: 414-428

Gianluca Demartini, Arjen P. de Vries, Tereza Iofciu, Jianhan Zhu: Overview of the INEX 2008 Entity Ranking Track. INEX 2008: 243-252