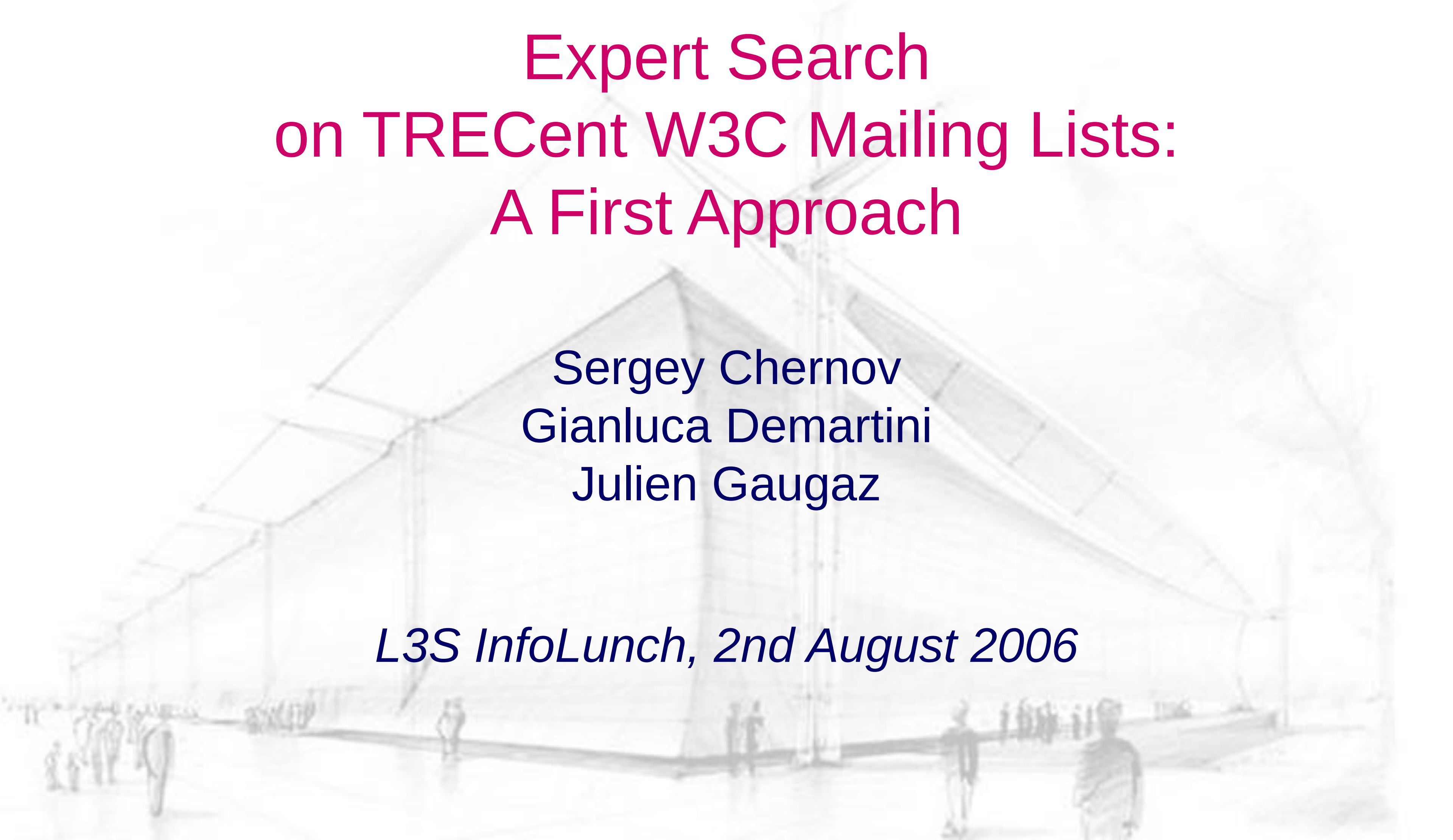
Expert Search on TRECent W3C Mailing Lists: A First Approach

Sergey Chernov
Gianluca Demartini
Julien Gaugaz

L3S InfoLunch, 2nd August 2006



Outline

- Introduction: TREC Enterprise Track 2006
- Expert Search – W3C test collection & Topics
- Our Approach:
  - Dummy Algorithm
  - More Clever Algorithms
- Learning from 2005 results
- Topics Specificity

Introduction

- **TREC:** Text REtrieval Conference standardizes evaluation in IR
- In 2005 the **Enterprise Track** started. Its goal is to study enterprise search: satisfying a user who is searching the data of an organization to complete some task
- One of the two tasks in this track is the **Expert Search**: find an expert on a given topic

Expert Search

- You are looking for a person or multiple people in your organization who are **experts on a subject**
- Reasons:
  - you need to **talk to someone** to get a starting point
  - you are trying to **assemble a project team**
- Expert search connects the documents to the people in the organization
- Think about collections for **social network analysis** and finding links between people

W3C Test Collection

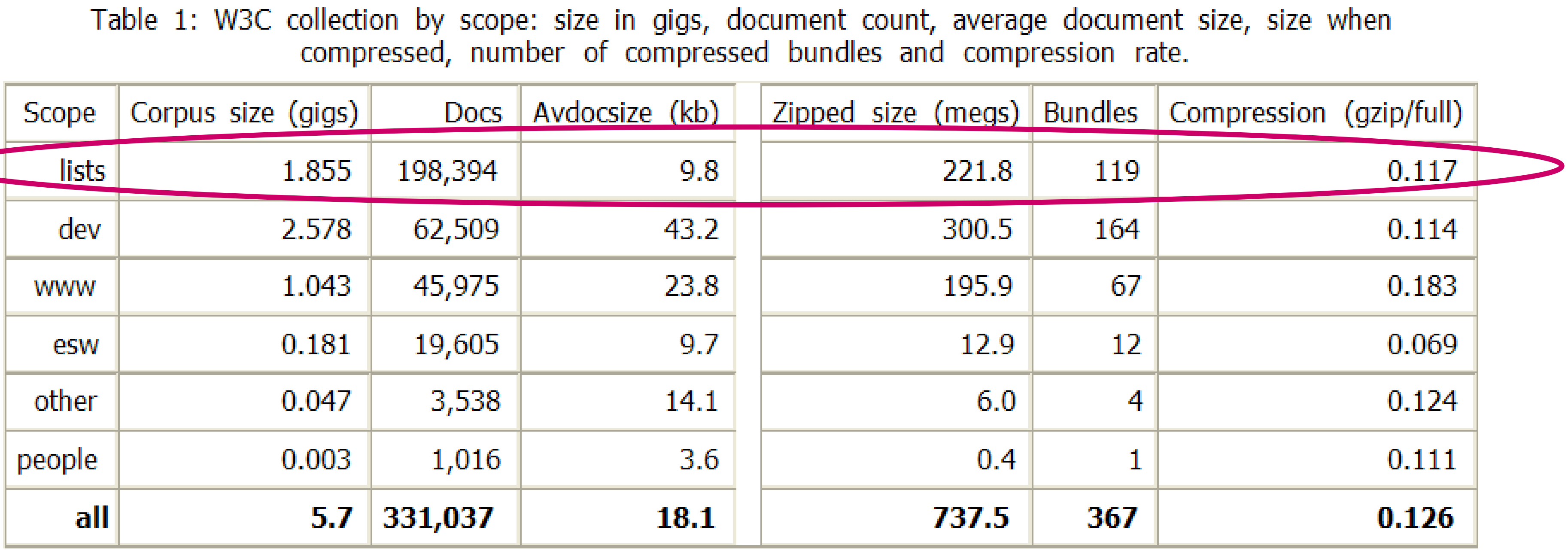
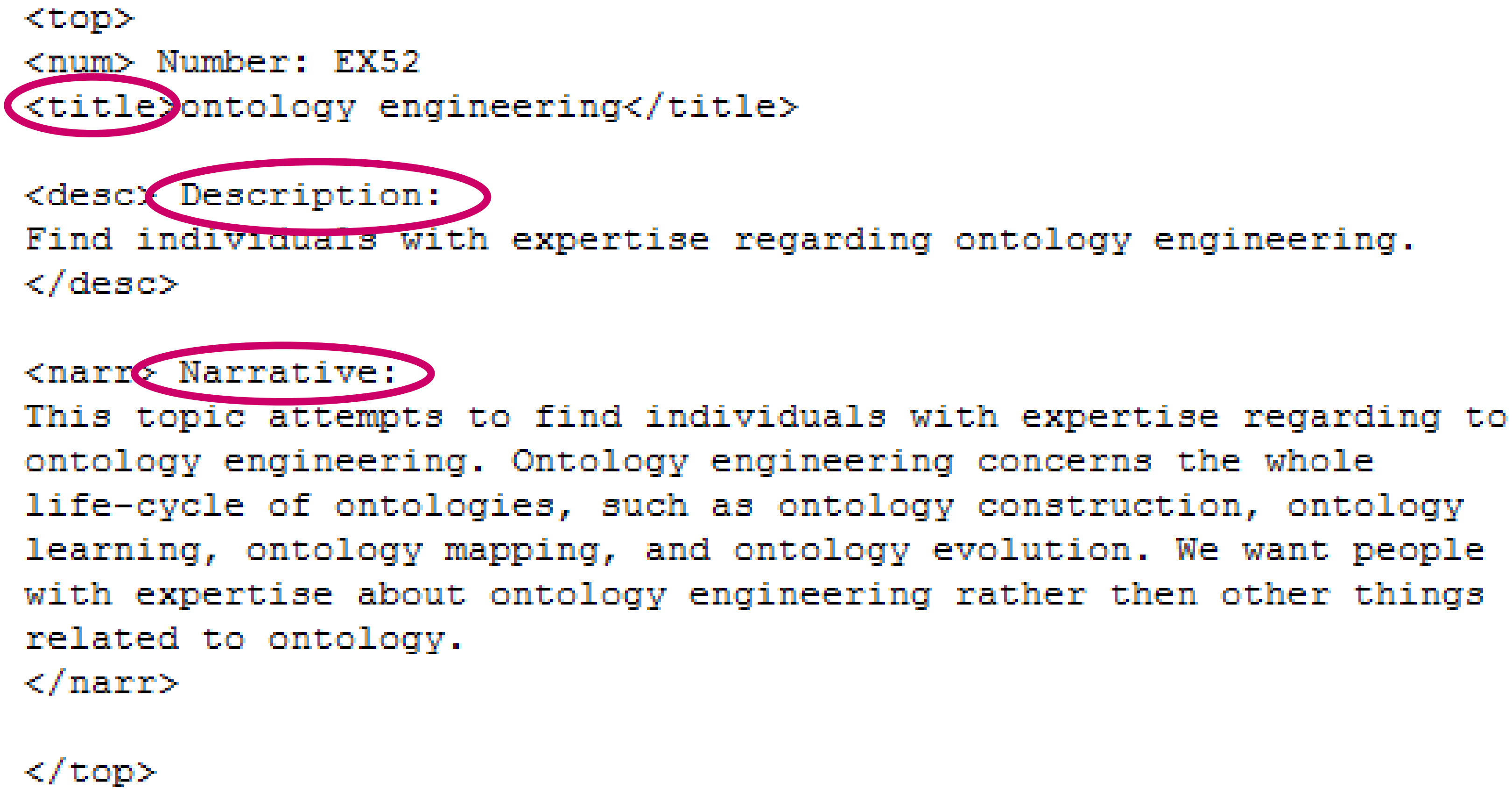


Table 1: W3C collection by scope: size in gigs, document count, average document size, size when compressed, number of compressed bundles and compression rate.

Scope	Corpus size (gigs)	Docs	Avdocsize (kb)	Zipped size (megs)	Bundles	Compression (gzip/full)
lists	1.855	198,394	9.8	221.8	119	0.117
dev	2.578	62,509	43.2	300.5	164	0.114
www	1.043	45,975	23.8	195.9	67	0.183
esw	0.181	19,605	9.7	12.9	12	0.069
other	0.047	3,538	14.1	6.0	4	0.124
people	0.003	1,016	3.6	0.4	1	0.111
all	**5.7**	**331,037**	**18.1**	**737.5**	**367**	**0.126**

Expert Search 2006: 55 Topics

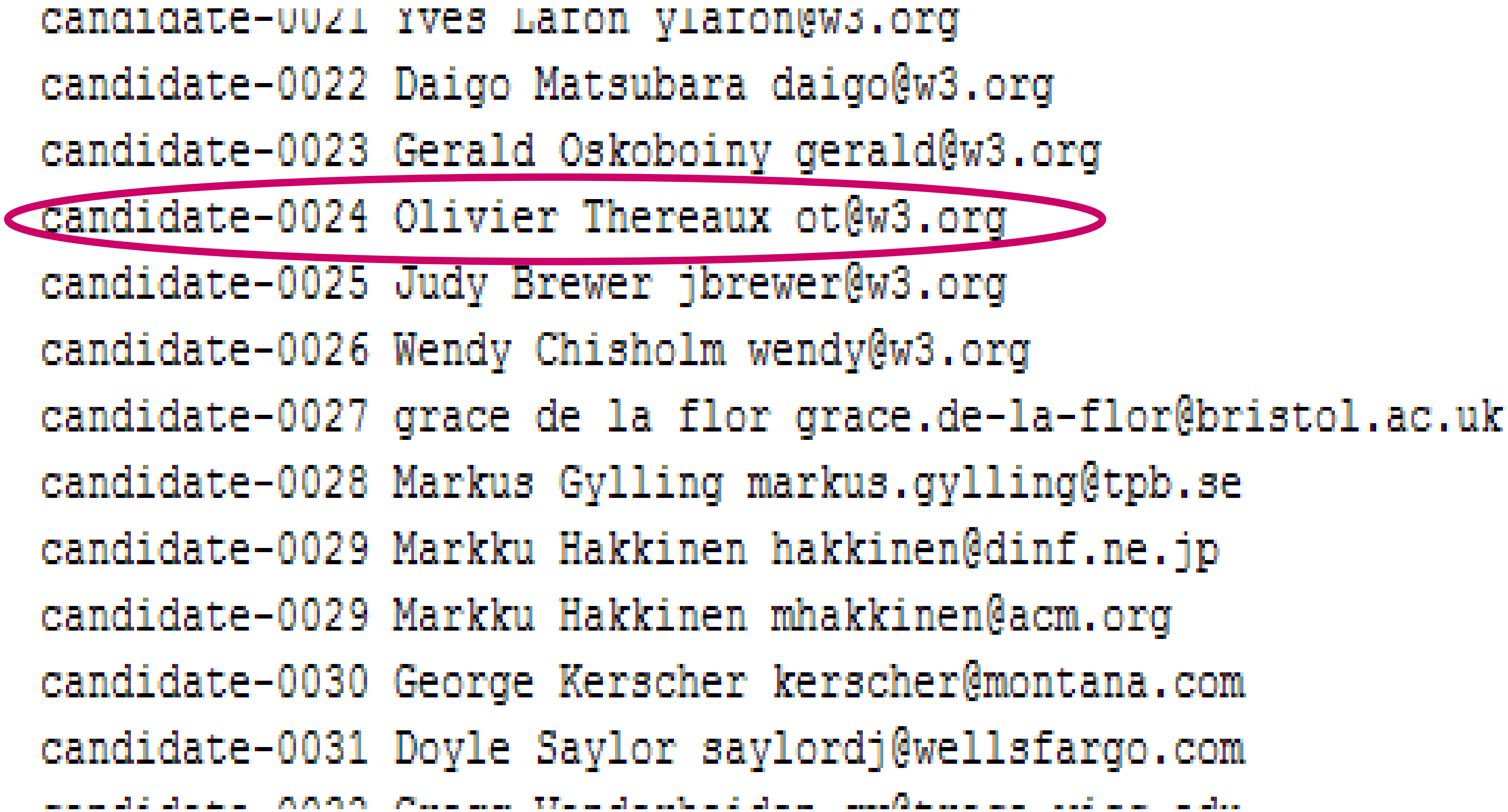
- 55 topics composed by title, description and narrative



```
<top>
<num> Number: EX52
<title>ontology engineering</title>

<desc> Description:
Find individuals with expertise regarding ontology engineering.
</desc>

<narr> Narrative:
This topic attempts to find individuals with expertise regarding to
ontology engineering. Ontology engineering concerns the whole
life-cycle of ontologies, such as ontology construction, ontology
learning, ontology mapping, and ontology evolution. We want people
with expertise about ontology engineering rather then other things
related to ontology.
</narr>

</top>
```

- In 2005 only title

Expert Search 2006: 1092 Candidates



```
candidate-0021 Yves Lafon ylafon@w3.org
candidate-0022 Daigo Matsubara daigo@w3.org
candidate-0023 Gerald Oskoboiny gerald@w3.org
candidate-0024 Olivier Thereaux ot@w3.org
candidate-0025 Judy Brewer jbrewer@w3.org
candidate-0026 Wendy Chisholm wendy@w3.org
candidate-0027 grace de la flor grace.de-la-flor@bristol.ac.uk
candidate-0028 Markus Gylling markus.gylling@tpb.se
candidate-0029 Markku Hakkinen hakkinen@dinf.ne.jp
candidate-0029 Markku Hakkinen mhakkinen@acm.org
candidate-0030 George Kerscher kerscher@montana.com
candidate-0031 Doyle Saylor saylordj@wellsfargo.com
```

A Fist Approach

- 2 weeks available: Only mailing list
- Mailing list cleaned to obtain an XML valid file
- Mailing list indexed with Lucene
- 4 different ways to find the experts on a given topic
  - 1 Dummy run: to have something to submit
  - 3 Clever runs:
    - Using **document score** threshold
    - Using **expert score** threshold
    - Using topic **specificity**

Run l3s1 (aka **Dummy** run)

- Requirement from TREC: only the Title part of the query is used
- Rank authors by #emails per author (in the relevant set)
- **expert score**: #emails
- Number of experts to be returned is set arbitrarily

Number of experts to retrieve = 5

Run l3s2 (aka **Documents score** run)

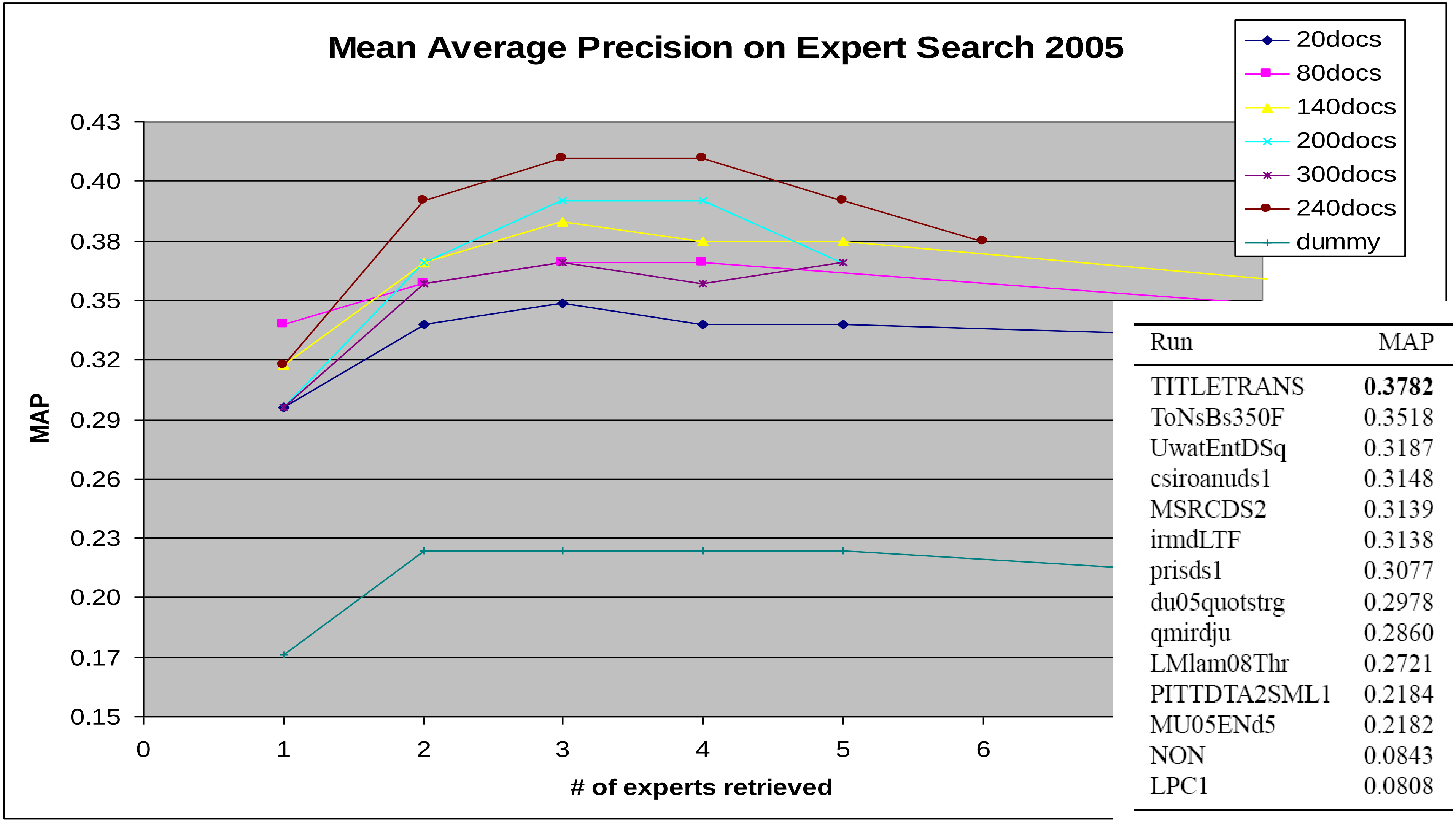
- **Documents score threshold** and **fixed number of expert**
- OR query
  - Title (weight 3.0)
  - Description (weight 2.0)
  - Narrative (weight 1.0)
- 80% documents are "relevant"
- Documents are relevant until sum over the first top-N documents below ***document threshold***
- **Assumption**: With low scores we need more docs to decide
- **Experts' score** is sum of scores of their emails (over the set of relevant emails)

Number of experts to retrieve = 5

Top-k documents considered relevant = 240

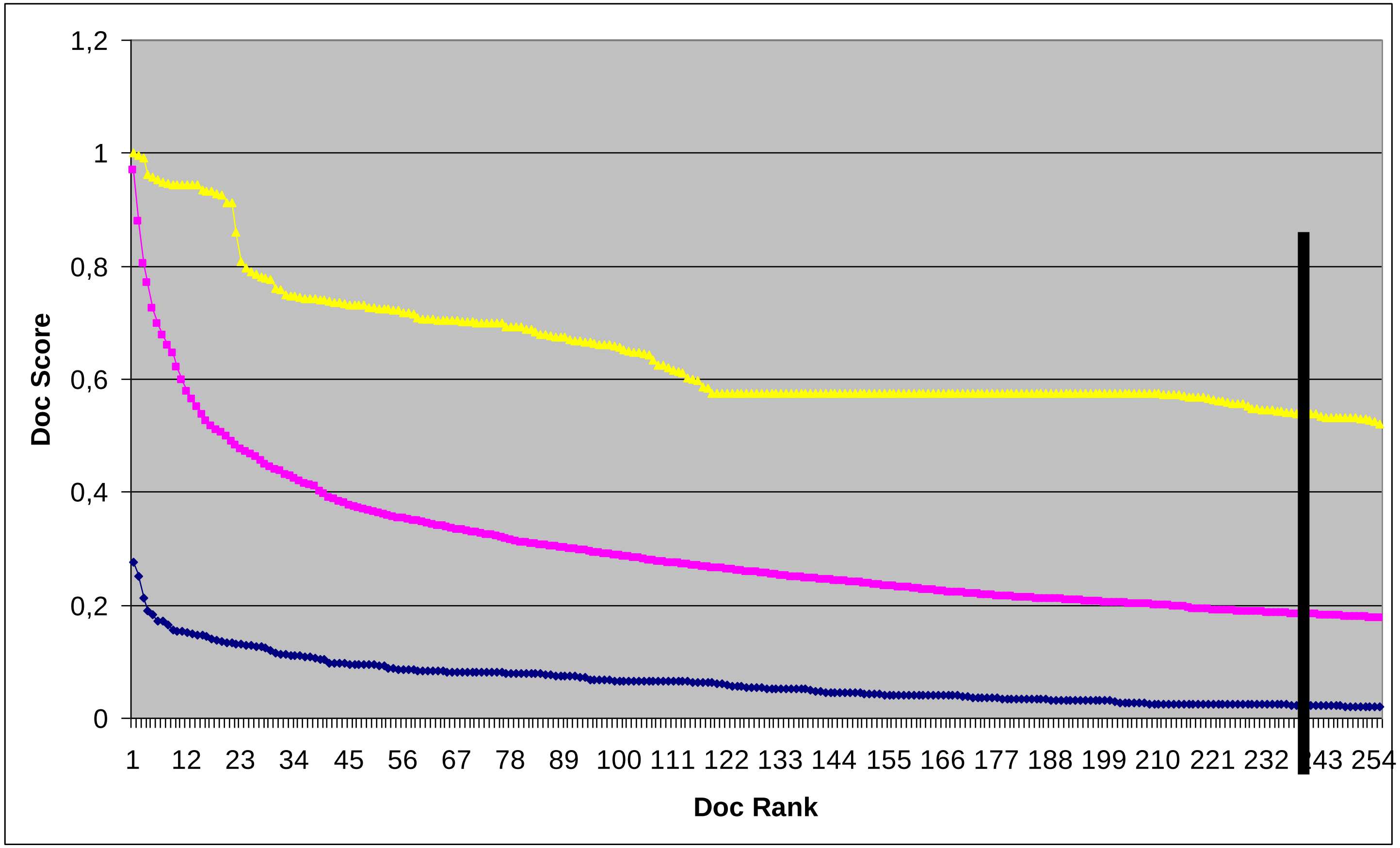
(sum of document RSV = 76.5)

Learning the parameters from the 2005 test collection



Run	MAP
TITLETRANS	**0.3782**
ToNsBs350F	0.3518
UwatEntDSq	0.3187
csiroanuds1	0.3148
MSRCDS2	0.3139
irndLTF	0.3138
prisds1	0.3077
du05quotstrg	0.2978
qmirdju	0.2860
LMlam08Thr	0.2721
PITTDTA2SML1	0.2184
MU05ENd5	0.2182
NON	0.0843
LPC1	0.0808

Doc Score on different rank position (2006)



Run l3s3 (aka **Expert score** run)

- **Documents score threshold** and **Expert score threshold**
- We retrieve all experts which score passes some threshold
- **Expert score**: score sum over all emails in the relevant set written by expert
- Doc threshold = fill the jar
- Expert threshold on expert score instead of fixed top-N

Expert score threshold = 1.2 = Avg expert score at rank 5

Top-k documents considered relevant = 240
(sum of document RSV = 76.5)

Run l3s4 (aka **Topics specificity** run)

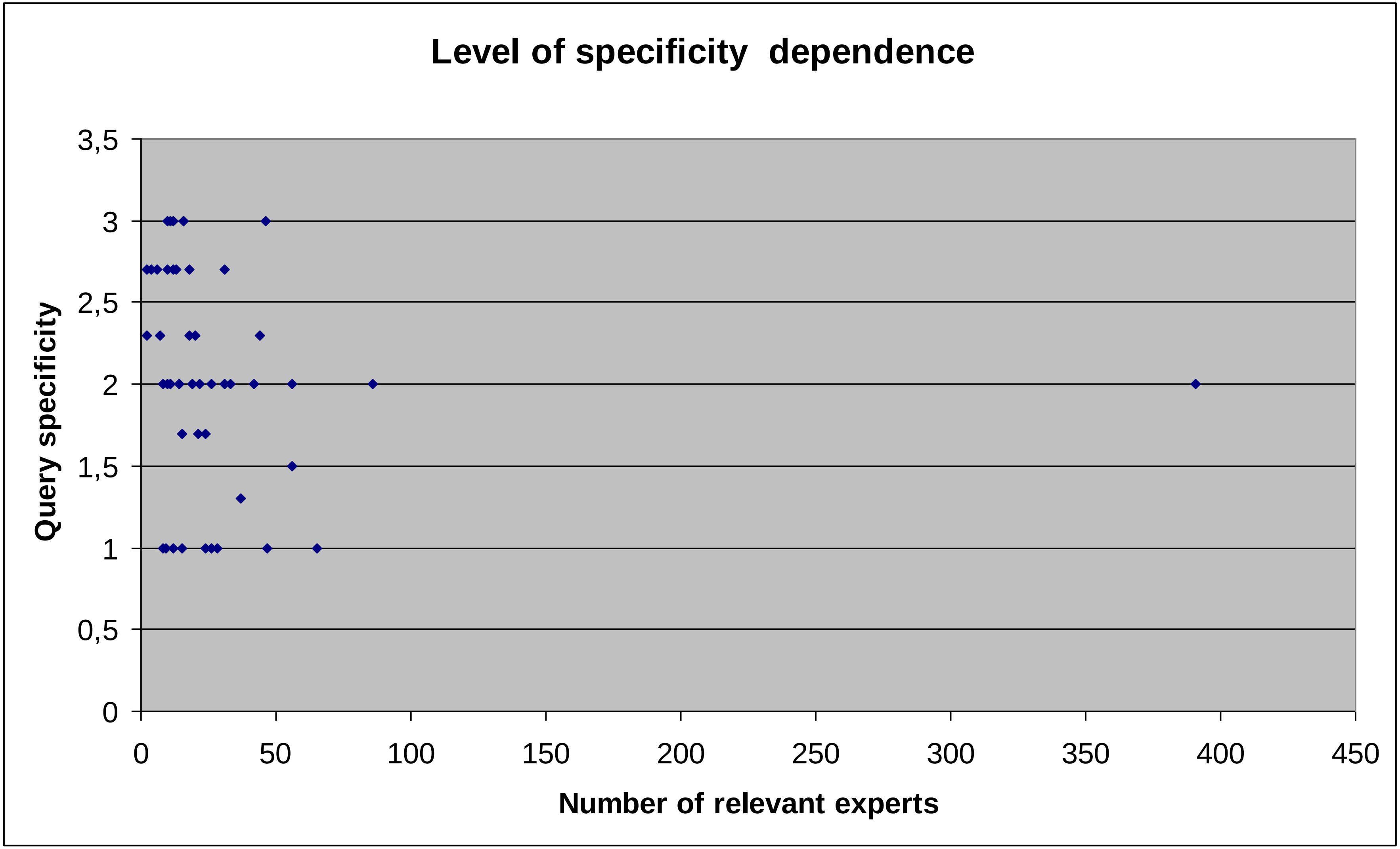
- **Documents score threshold** and **different Expert score thresholds**
- **Expert threshold:** sum of scores of retrieved relevant documents written by an expert, multiplied by the **topic specificity** value
- Topic Specificity value

  0.5 <= TS <= 1.5
  - 0.5 **general** (many experts)
  - 1.5 **very specific** (few experts)

Each query gets its specificity level as a number from 0.5 to 1.5
Expert score threshold = 1.2*specificity

Top-k documents considered relevant = 240
(sum of document RSV = 76.5)

On topic Specificity (evaluation on 2005 test collection)



Future Work

- Expert Search in Beagle ++ ?
- Expert Search using PLSA ?

Conclusions

- At least one run (l3s2) has good results on the 2005 collection
- Topic Specificity seems to be not correlated with the number of experts (lack of definition…)

TRECent Expert Search 2006: Important dates

30 July: Discussion search and Expert search runs due

Mid August to Mid September: Relevance judging for expert search

September: Results available

October: TREC notebook papers due

14-17 November: TREC

Thanks for your attention!

Q&A