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>
<thead>
<tr>
<th>Scope</th>
<th>Corpus size (gigs)</th>
<th>Docs</th>
<th>Avdocsize (kb)</th>
<th>Zipped size (megs)</th>
<th>Bundles</th>
<th>Compression (gzip/full)</th>
</tr>
</thead>
<tbody>
<tr>
<td>lists</td>
<td>1.855</td>
<td>198,394</td>
<td>9.8</td>
<td>221.8</td>
<td>119</td>
<td>0.117</td>
</tr>
<tr>
<td>dev</td>
<td>2.578</td>
<td>62,509</td>
<td>43.2</td>
<td>300.5</td>
<td>164</td>
<td>0.114</td>
</tr>
<tr>
<td>www</td>
<td>1.043</td>
<td>45,975</td>
<td>23.8</td>
<td>195.9</td>
<td>67</td>
<td>0.183</td>
</tr>
<tr>
<td>esw</td>
<td>0.181</td>
<td>19,605</td>
<td>9.7</td>
<td>12.9</td>
<td>12</td>
<td>0.069</td>
</tr>
<tr>
<td>other</td>
<td>0.047</td>
<td>3,538</td>
<td>14.1</td>
<td>6.0</td>
<td>4</td>
<td>0.124</td>
</tr>
<tr>
<td>people</td>
<td>0.003</td>
<td>1,016</td>
<td>3.6</td>
<td>0.4</td>
<td>1</td>
<td>0.111</td>
</tr>
<tr>
<td>all</td>
<td>5.7</td>
<td>331,037</td>
<td>18.1</td>
<td>737.5</td>
<td>367</td>
<td>0.126</td>
</tr>
</tbody>
</table>
Expert Search 2006: 55 Topics

- 55 topics composed by title, description and narrative

```xml
<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 rives laron riveslaron@w3.org

candidate-0022 Daigo Matsubara daigo@w3.org

candidate-0023 Gerald Oskoboiny gerald@w3.org

candidate-0024 Olivier Thereaux othe@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

candidate-0032 Svante Hedingheide sv@bubbe.wisc.edu
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

Mean Average Precision on Expert Search 2005

<table>
<thead>
<tr>
<th>Run</th>
<th>MAP</th>
</tr>
</thead>
<tbody>
<tr>
<td>TITLETRANS</td>
<td>0.3782</td>
</tr>
<tr>
<td>ToNsBs350F</td>
<td>0.3518</td>
</tr>
<tr>
<td>UwatEntDSq</td>
<td>0.3187</td>
</tr>
<tr>
<td>csiroanuds1</td>
<td>0.3148</td>
</tr>
<tr>
<td>MSRCDS2</td>
<td>0.3139</td>
</tr>
<tr>
<td>irmdLTF</td>
<td>0.3138</td>
</tr>
<tr>
<td>prisd1</td>
<td>0.3077</td>
</tr>
<tr>
<td>du05quotstrg</td>
<td>0.2978</td>
</tr>
<tr>
<td>qmirdju</td>
<td>0.2860</td>
</tr>
<tr>
<td>LMLam08Thr</td>
<td>0.2721</td>
</tr>
<tr>
<td>PITTDTA2SML1</td>
<td>0.2184</td>
</tr>
<tr>
<td>MU05End5</td>
<td>0.2182</td>
</tr>
<tr>
<td>NON</td>
<td>0.0843</td>
</tr>
<tr>
<td>LPC1</td>
<td>0.0808</td>
</tr>
</tbody>
</table>
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**
  - \(0.5 \leq TS \leq 1.5\)
    - 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