Improving the Estimation of Relevance Models Using Large External Corpora

Don Metzler

Joint work with Fernando Diaz

Motivation

- Performance of pseudo-relevance feedback (PRF) depends critically on ability to find relevant material in collection
- What if there is little or no relevant material in collection?
- Idea: expand against target collection plus one or more external collections

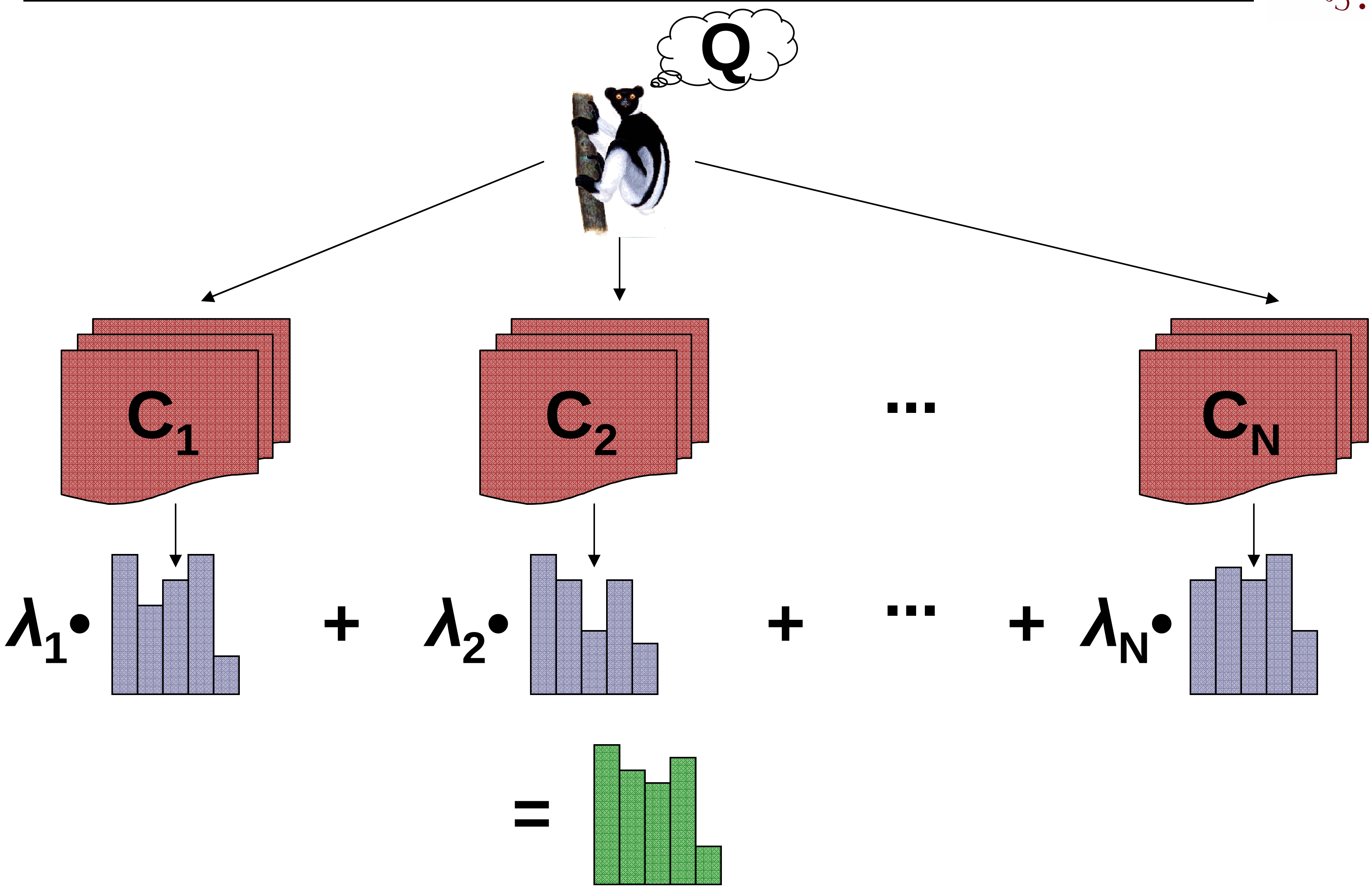
Past Work

- TREC 6
  - “increasing the size of the database increases the likelihood of finding good expansion concepts” (Allan et. al., 1997)
  - “it is quite clear that ‘blind’ query modification is beneficial provided that a large enough database is available” (Walker et. al., 1997)
- Local context analysis (LCA)
  - LCA on larger, external collection led to 12% increase in 11-point average precision (Xu and Croft, 2000)
- TREC Robust Track
  - Web expansion

Our Goals

- Formalize multi-collection PRF
- Perform more detailed experiments
  - Does external expansion help for ad hoc retrieval on a web collection?
  - Is web expansion always the best or can we get by with expanding using something smaller?
  - How does external expansion compare to true feedback?
- Develop a better understanding of PRF
  - When and why does external expansion work?

Model



Model

- Compute relevance model over each collection using top *M* documents
- Mix models together
- Sample top *K* terms from combined model and form expanded Indri query
  - **`#weight( P(w`$_1$`|R) w`$_1$` … P(w`$_K$`|R) w`$_K$` )`**
- Mix original query with expanded query terms by forming Indri query
  - **`#weight( λ Q`$_{ORIG}$` (1–λ) Q`$_{EXPANDED}$` )`**

Experimental Setup

- Experiments done using Indri
- Query-time stopping
- Krovetz stemming
- Evaluation uses 10-fold cross validation

	Collection	*Documents*	*Topics*
Target	trec12	469,949	150
	robust	472,525	250
	wt10g	1,692,096	100
External	bignews	6,422,629	-
	gov2	25,205,179	-
	web	19,200,000,000	-

Results

			BIGNEWS		GOV2		WEB	
	QL	RM3	External	Mixture	External	Mixture	External	Mixture
trec12	0.2502	0.3201	0.3204	**0.3319**	*0.2709*	0.3215	*0.3092*	**0.3324**
robust	0.2649	0.3214	**0.3501**	**0.353**	*0.2748*	0.3207	0.3301	**0.3352**
wt10g	0.1982	0.203	**0.2256**	**0.2331**	0.1999	0.1958	**0.2452**	**0.2429**

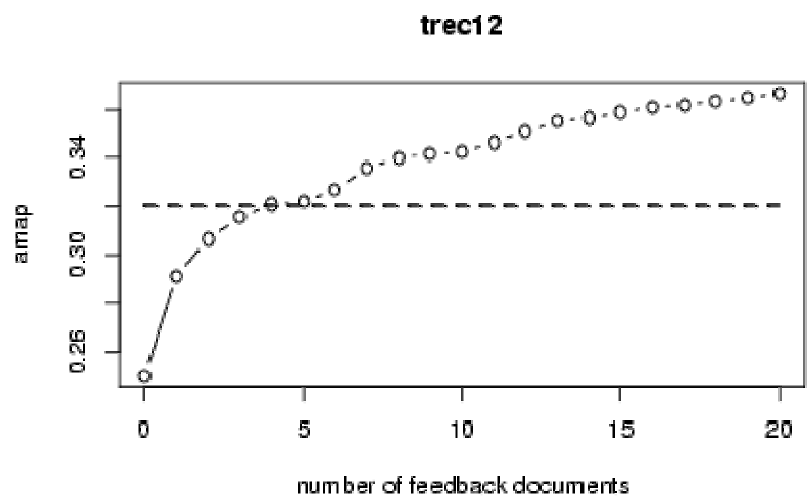
- **QL**: query likelihood
- **RM3**: original query + target RM
- **External**: original query + external RM
- **Mixture**: original query + mixture of target and external RM

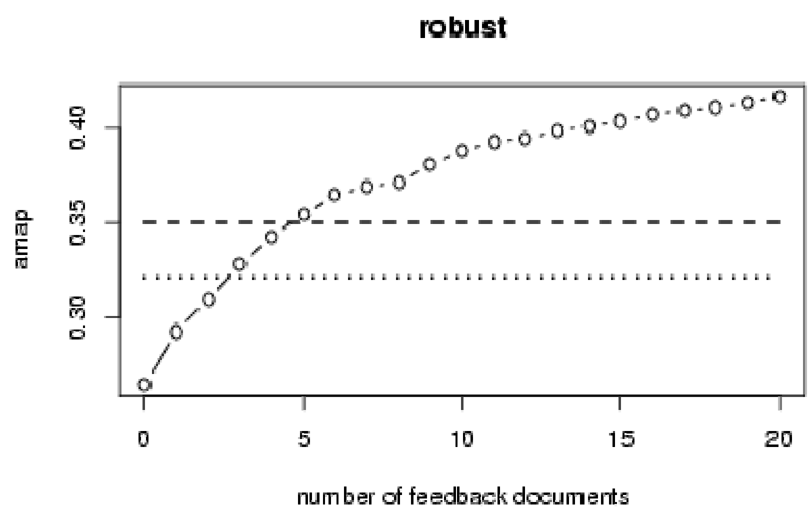
- **Bold**: statistically significantly better than RM3
- *Italics*: statistically significantly worse than RM3

Pseudo vs. True Feedback

- Pseudo-relevance feedback using external collections yields strong effectiveness
- How does it compare to true feedback?
- Experimental Setup
  - Simulate feedback on top *K* documents
  - Construct ‘true’ relevance model from documents judged relevant
  - Formulate query as mixture of original query and top terms from this model

Feedback Results





- trec12
  - both types of expansion equivalent to feedback on 5 documents
- robust
  - target expansion equivalent to feedback on 2 documents
  - external expansion equivalent to feedback on 5 documents

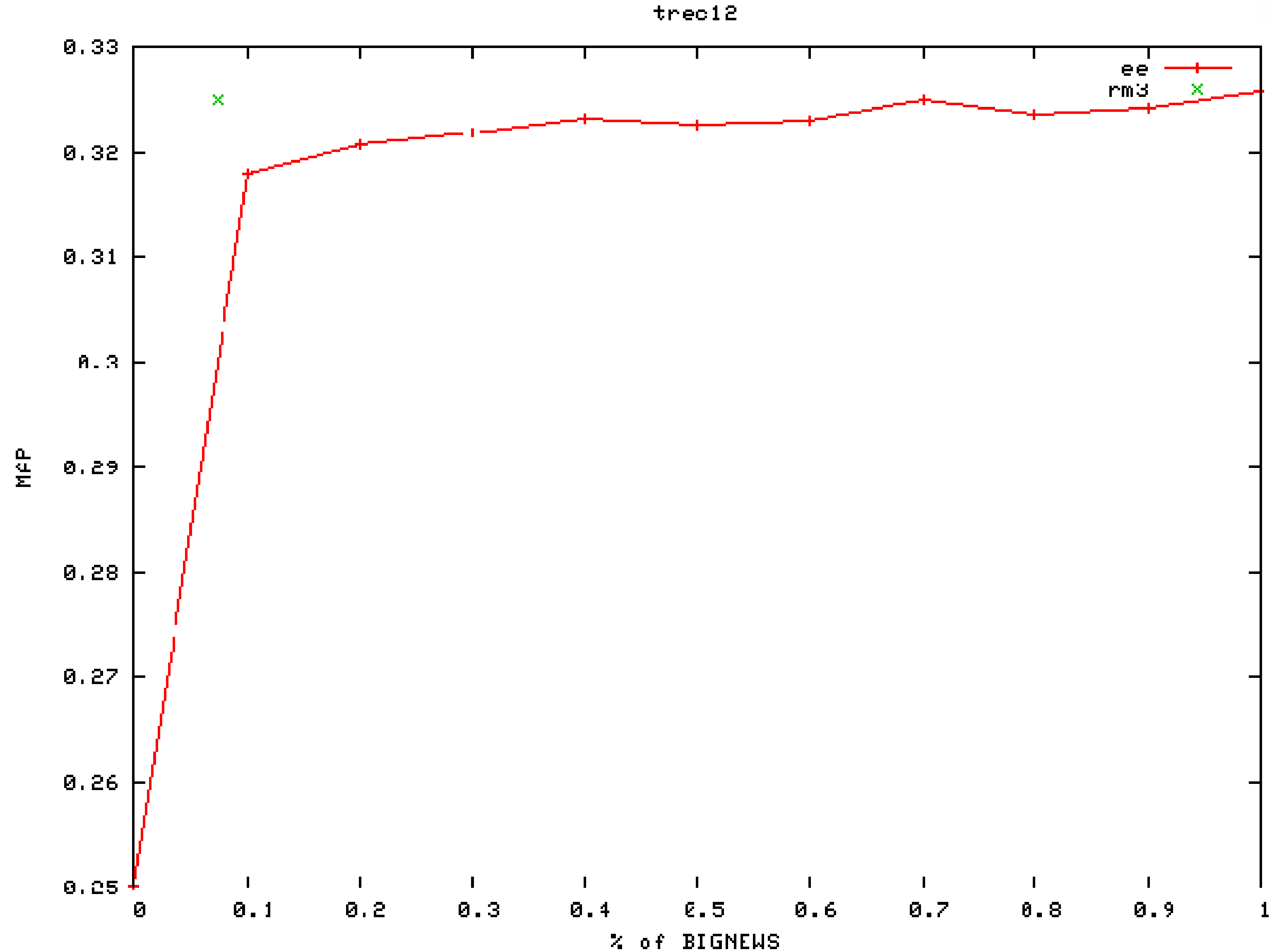
Legend

- dotted line = PRF using target collection
- dashed line = PRF using external collection

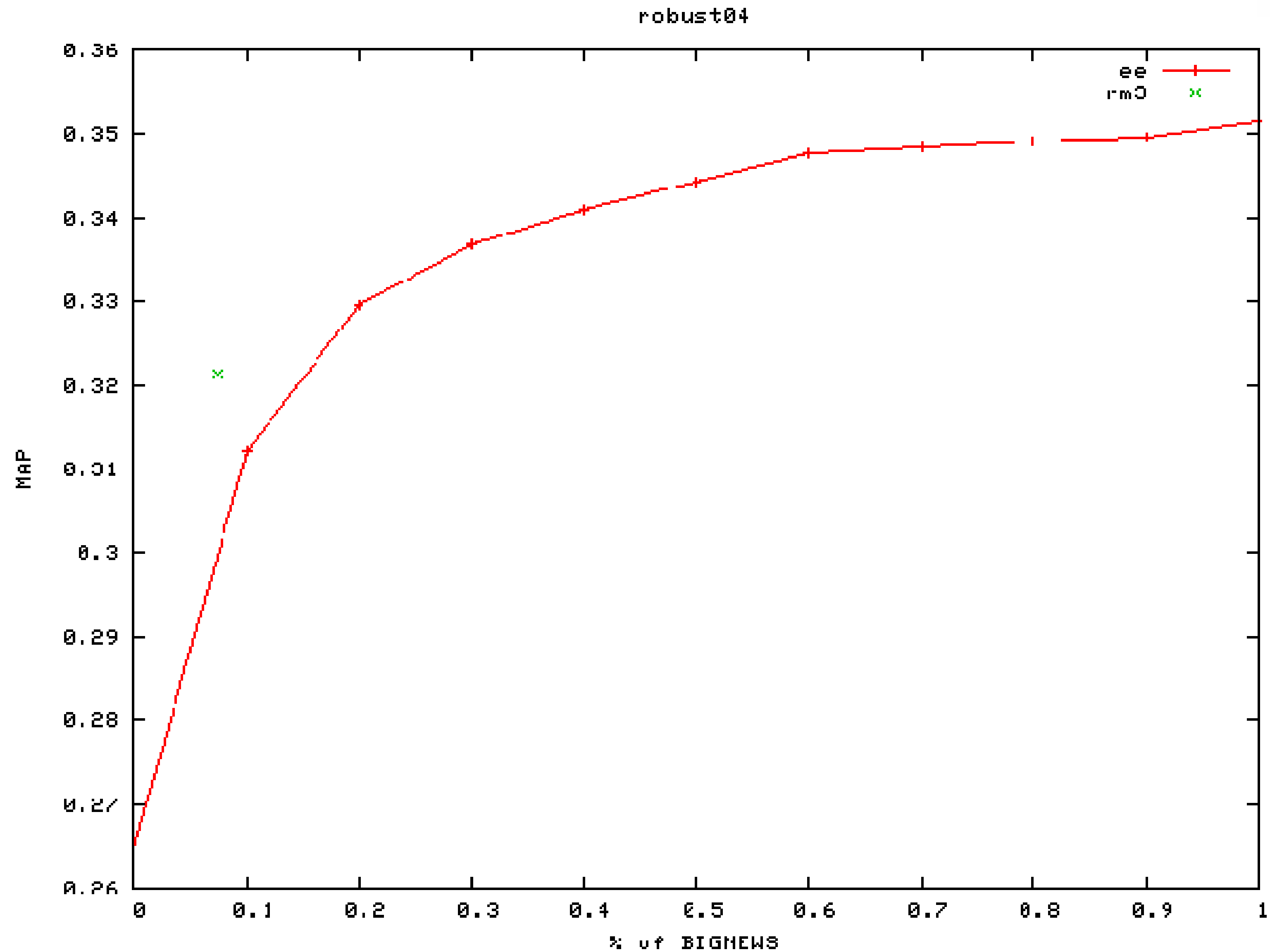
Collection Size Effects

- What is the relationship between external collection size and effectiveness?
- How large must the external collection be before we see diminishing returns?
- Experimental Setup
  - Randomly subsample documents from BIGNEWS
  - Use sampled collection for external expansion
  - Plot MAP vs. external collection size

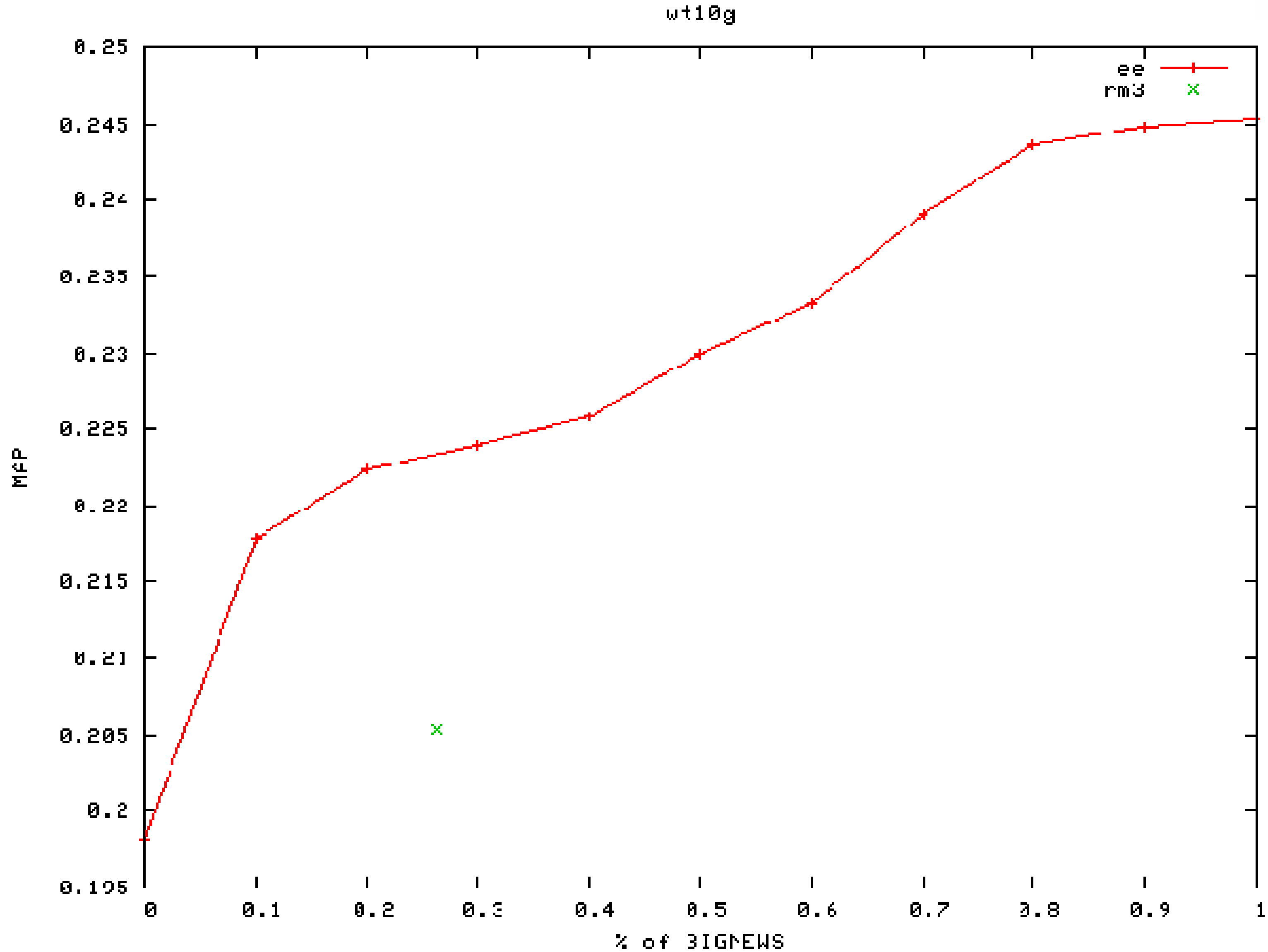
Size Experiments



Size Experiments



Size Experiments



Concept Density

- Can we predict the best collection to use for expansion?
- Collection size not only important factor
  - Expanding by GOV2 was not helpful
- Topic coverage important
  - Estimated RM likely to be poor for topics with poor coverage
  - *Concept density* measures how densely represented query concepts are in a collection

Concept Density

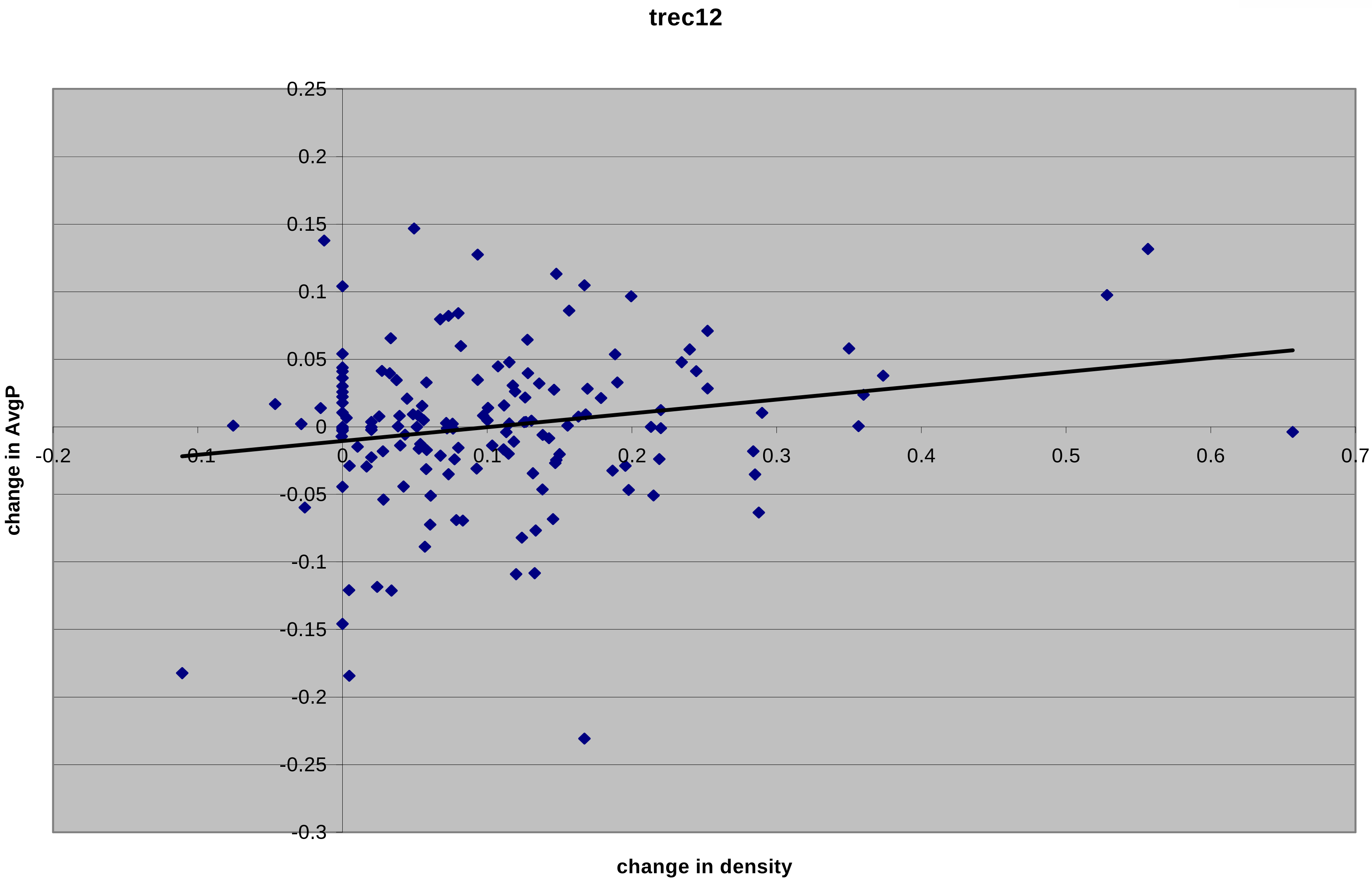
- Basic steps:
  - Extract concepts from query
  - Compute concept density for each collection
  - (optional) Construct RM from collection with highest concept density
- Extracting concepts
  - Use 'dependence model' concepts
- Computing density
  - Proportion of top *K* documents that contain the concept

Concept Density Example

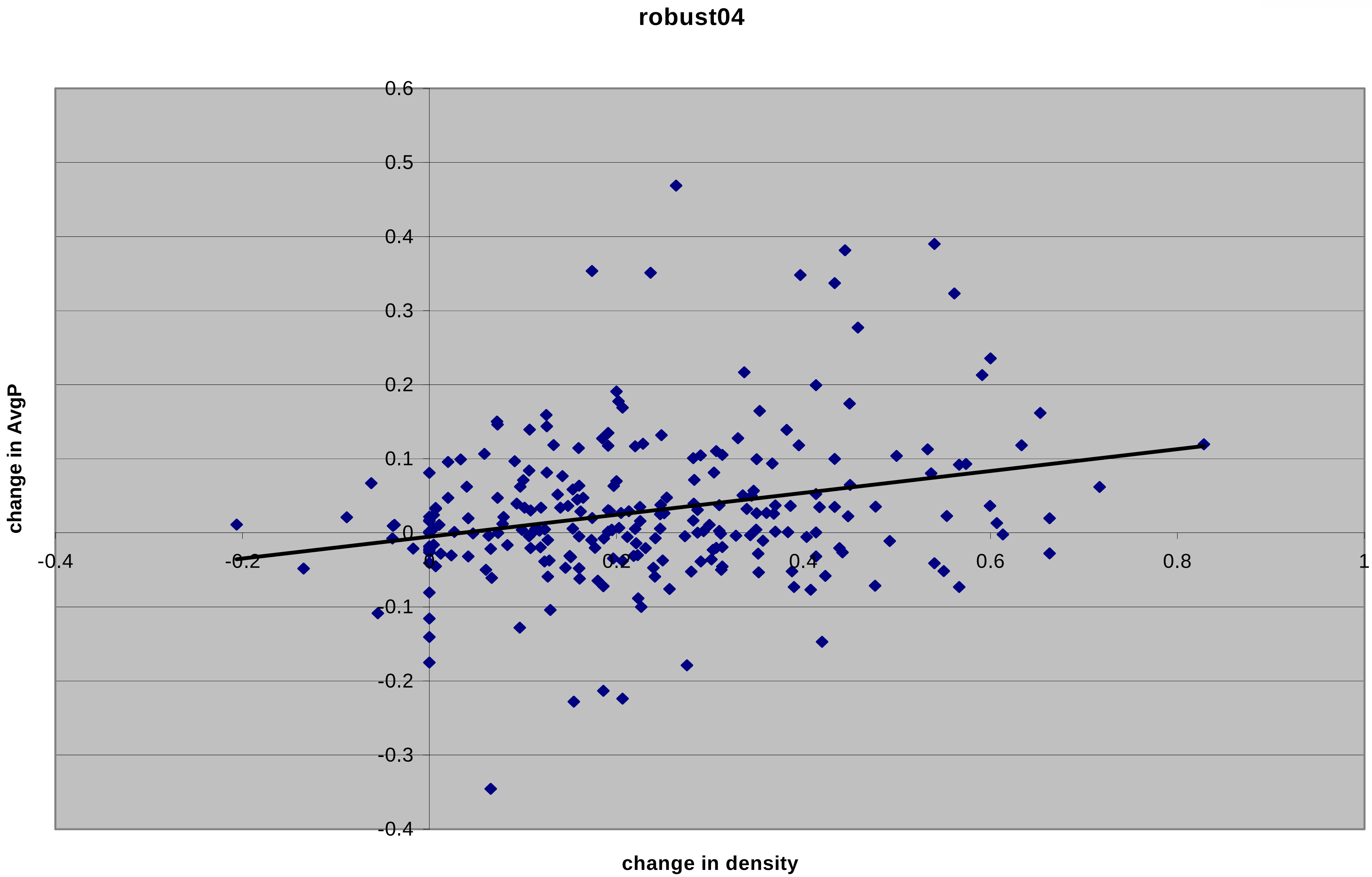
Query: olympics training swimming

<table>
<tr><th>Concept</th><th colspan="2">robust density</th><th colspan="2">bignews density</th></tr>
<tr><td>olympics</td><td>1</td><td rowspan="3">0.95</td><td>1</td><td rowspan="3">1</td></tr>
<tr><td>training</td><td>0.88</td><td>1</td></tr>
<tr><td>swimming</td><td>0.98</td><td>1</td></tr>
<tr><td>#1(training swimming)</td><td>0.02</td><td rowspan="3">0.04</td><td>0.02</td><td rowspan="3">0.09</td></tr>
<tr><td>#1(olympics training)</td><td>0.1</td><td>0.24</td></tr>
<tr><td>#1(olympics training swimming)</td><td>0</td><td>0</td></tr>
<tr><td>#uw8(training swimming)</td><td>0.36</td><td rowspan="4">0.31</td><td>0.94</td><td rowspan="4">0.89</td></tr>
<tr><td>#uw8(olympics swimming)</td><td>0.62</td><td>0.98</td></tr>
<tr><td>#uw8(olympics training)</td><td>0.22</td><td>0.88</td></tr>
<tr><td>#uw12(olympics training swimming)</td><td>0.04</td><td>0.74</td></tr>
<tr><td>overall density</td><td colspan="2">0.43</td><td colspan="2">0.66</td></tr>
</table>

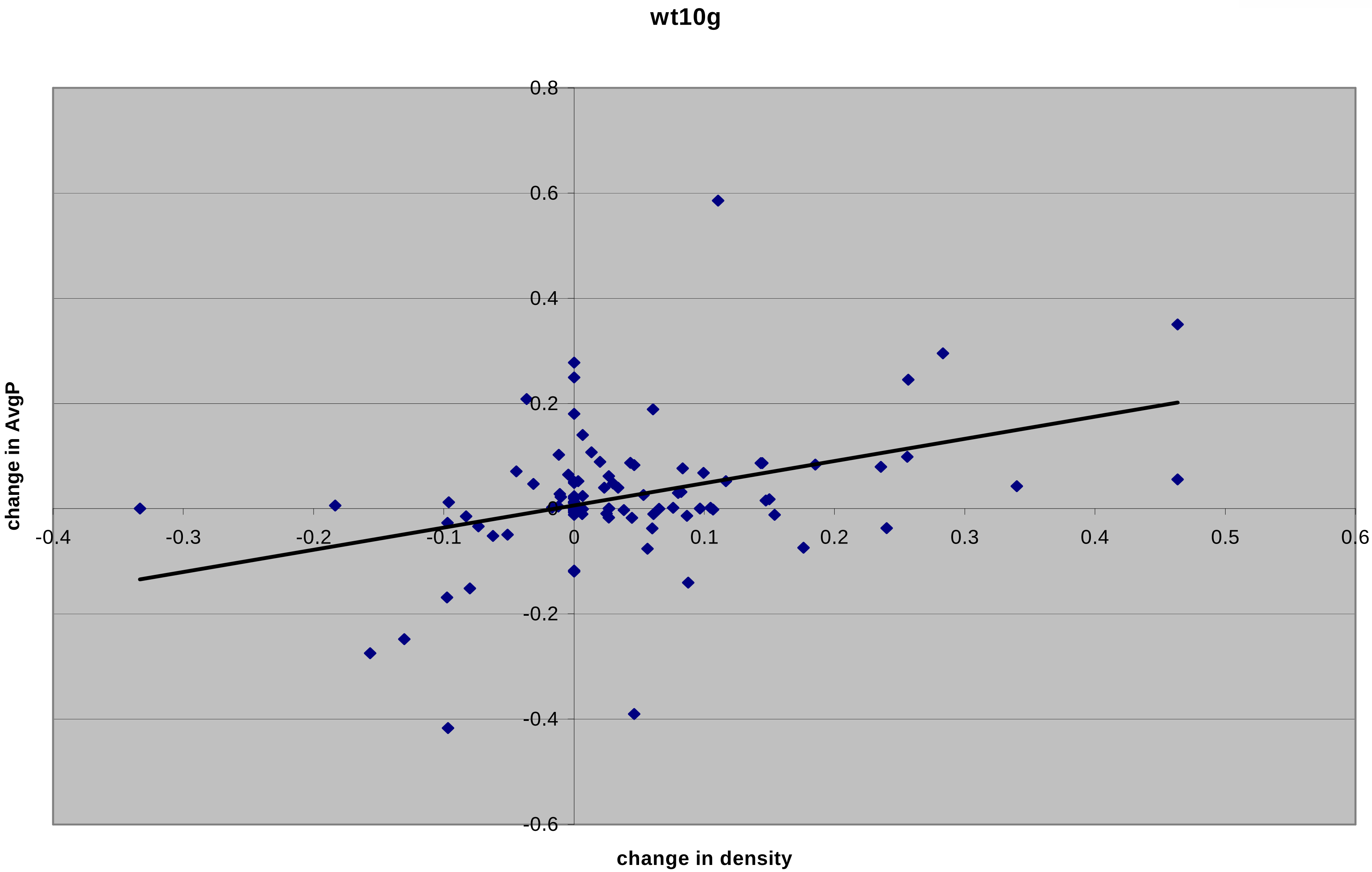
Concept Density Results



Concept Density Results



Concept Density Results



Conclusions

- Contributions
  - Formal method for performing PRF using one or more external collections
  - Comparison of PRF using large news collection and the web as external resources
  - Used concept density to developing a better understanding of why PRF works/fails
- Future work
  - Use Wikipedia as external resource
  - Further exploration of concept density