Web science

The Importance of Anchor Text for Ad Hoc Search Revisited

Matsuev Egor

2019

Overview

Problem: no clear evidence of ad hoc retrieval effectiveness using web links
It was thought that it is due to either:

- small dataset
- small inter-server links density

Conclusions:

1. link density has little impact on effectiveness, mainly depends on size of dataset
2. compared different approaches for information retrieval
3. propagated anchor text outperforms full-text retrieval
   in terms of early precision, and in combination with it, gives an improvement in overall precision.

What is the subject?

TREC- Text Retrieval Conference

- 1999 WT10g+other database (no result, big difference user-ad hoc)
- WT-dense subset of WT10g with bigger inter-sever link density (better results)
- 2004-2006 GOV2 25 mln docs (no good results for ad-hoc search)
- 2009 ClueWeb09

Factors

- **Intra-links:** links within the same site are often navigational links, with anchor terms such as ‘click’, ‘here’ and ‘next’
- **inter-links:** links between sites: they are more meaningful
- **bigger database** more incoming links, more relevant documents to a given topic.

adhoc search the task is to find pages with relevant text no matter their popularity

Initial experiments

ClueWeb09 category B
two indexes (Indri tool):

- a full-text index containing only the document text
- an anchor text (underlined links) index containing only the propagated anchor text.

1.18 billion links between over 50 million pages
Documents are scored using the document length
normalised mixture run $Score_{mix}(d) = 0.7 \cdot Score_{full}(d) + 0.3 \cdot Score_{anchor}(d)$
In-degree rate by the number of incoming links

Results

- statAP and MPC evaluate runs
- degree-based link are good for separating pages, in this dataset pages are of high quality

anchor is more sensitive to the topical context than in-degree

Table 1: Results for the 2009 Adhoc Task. Significance tests are with respect to the full text run, confidence levels are 0.95 (°), 0.99 (°) and 0.999 (•)

Run	Full collection		No Wikipedia	
	statMAP	MPC(30)	statMAP	MPC(30)
Text	0.1442	0.3079	0.1038	0.2557
Anchor	0.0567	**0.5558**	0.0617	0.4289
Mix	**0.1643**°	0.4812°	0.1213	0.4773
In-degree	0.0823	0.1876	0.0592	0.1258
Text · In-degree	0.1098	0.2694	0.0746	0.2059
UDWAxQEWeb	0.1999	0.5010	–	–
uogTrdphCEwP	0.2072	0.4966	–	–
ICTNETADRun4	0.1746	0.4368	–	–

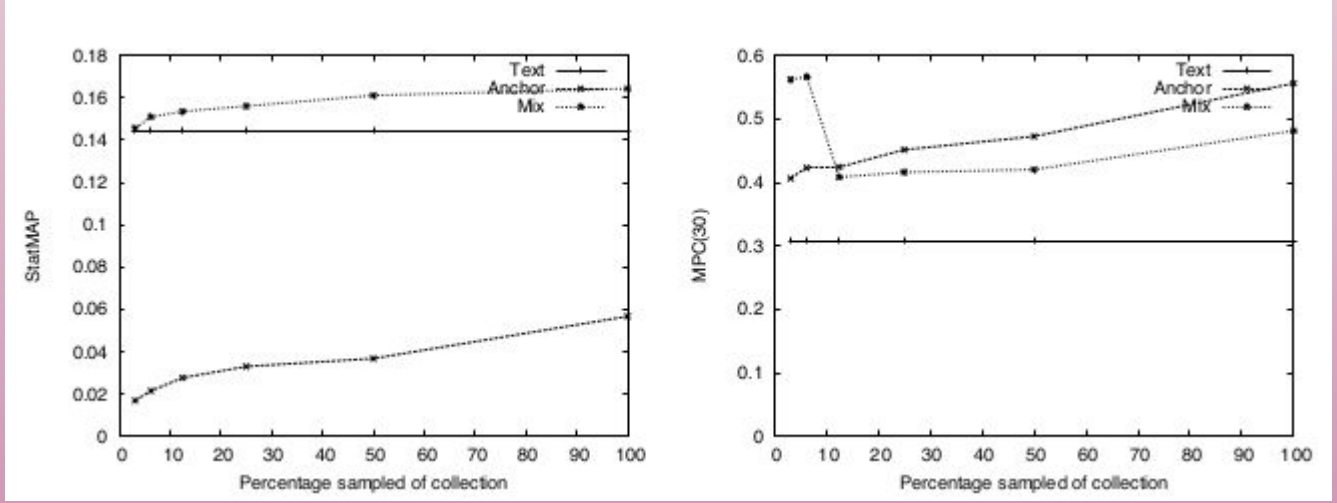
the impact of link density

- both the number of inter- and intra-server links decrease, linear to the sample size.
- The impact of link density is small.
- The highest quality pages have so many incoming links that they are robust against link sampling
- It plays a role at low densities, but its impact stabilises quickly

Table 2: Impact of link filtering on the percentage of pages with anchor text

	All pages			Relevant pages		
Percent	Inter	Intra	All	Inter	Intra	All
100.000	15.30	70.26	75.43	25.54	74.46	80.96
50.000	11.41	56.35	61.51	21.04	64.84	71.96
25.000	8.24	43.79	48.36	17.14	54.87	61.97
12.500	5.78	33.06	36.75	13.77	44.15	50.75
6.250	3.94	24.17	26.96	10.94	35.81	41.80
3.125	2.61	17.00	19.00	8.35	28.41	33.33

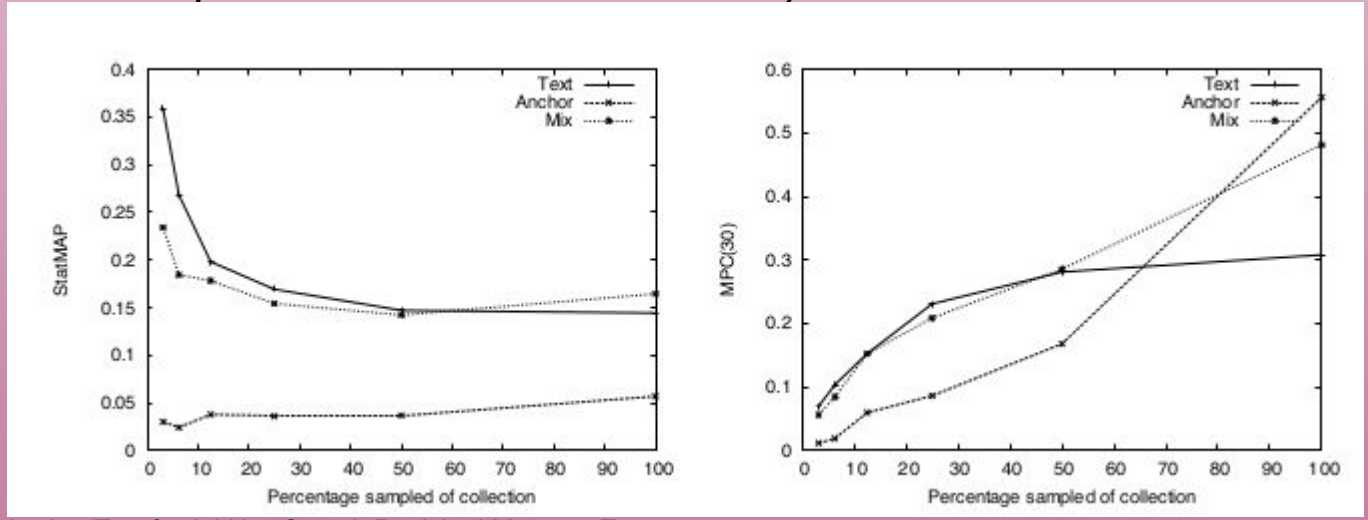
The impact of link density



with smaller samples, the MPC(30) scores of the anchor text run stay well above the Text score. The Mix run scores better at statMAP with even the smallest samples of links, indicating that even very few links can improve the Text run the highest quality pages have so many incoming links that they are robust against link sampling

The Impact of Collection Size

At each step, the ratio of all pages and relevant pages is roughly the same page sampling has a similar impact on the coverage of inter-server links as link sampling, but a very different impact on the coverage of intra-server links. The anchor text precision is more affected by collection size.



Thank you for your attention!