Measuring Diversity of a Domain-Specific Crawl

CICLing 2015

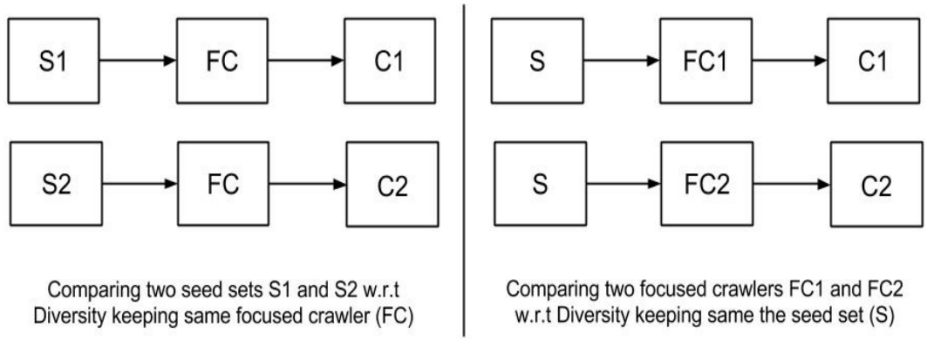
P Nikhil Priyatam
Krish Perumal
Vasudeva Varma

Search and Information Extraction Lab
International Institute of Information Technology, Hyderabad, India

IIIT-H

Motivation

- People use domain specific search engines for finding good results.
- Zero results for an “in-domain” query is a very bad situation for a domain specific search engine. Clearly, diversity of crawled content is crucial for a SE.
- Typically, domain specific search engines rely on focused crawlers to crawl content.
- The performance of a focused crawler in-turn depends on the crawling strategy and the seed set it uses.

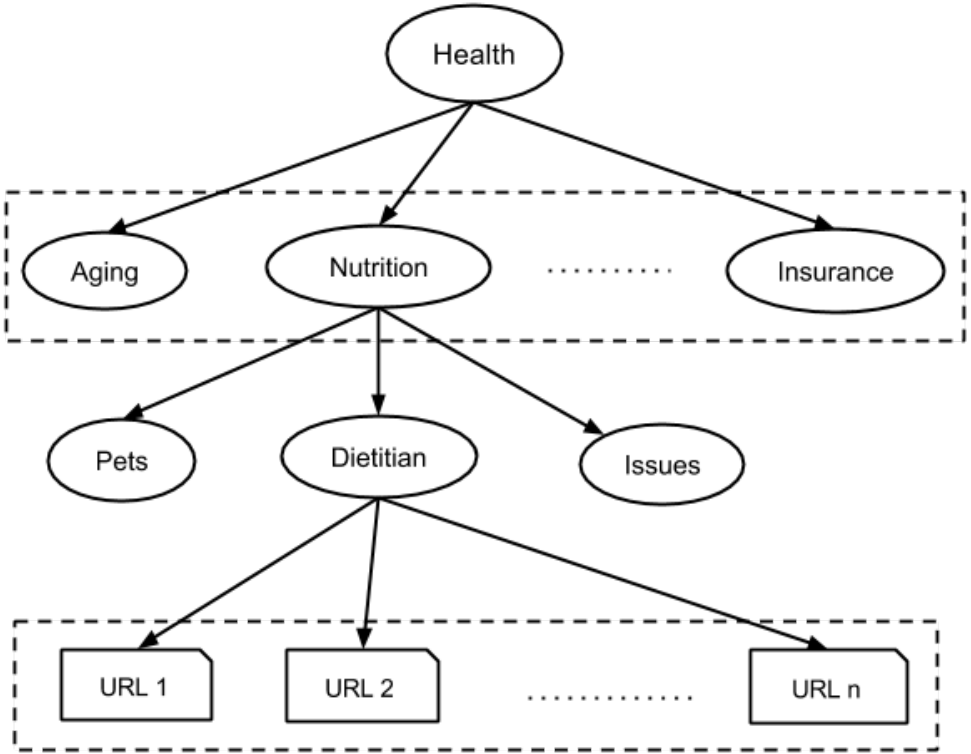


Metrics to Measure Crawl Diversity

We tried 4 different metrics to measure diversity of a domain specific crawl:

- Semantic Similarity Based Metric
- Dispersion Based Metric
- Topic Modeling Based Metric
- Cosine Similarity Based Metric

Experimental Setup



- We generate a **more diverse** and a **less diverse** seed set for a domain using ODP URLs.
- The evaluation metric is assumed to capture diversity if Diversity score of **More diverse** crawl is **strictly greater than** diversity of **Less diverse** crawl.

Conclusion

- Experiments show that, cosine similarity and dispersion based metrics outperform semantic distance and topic modeling based metrics.

- To the best of our knowledge, we are the first ones to evaluate a domain specific crawl w.r.t Diversity.