Seed Selection for Domain-Specific Search

Pattisapu Nikhil Priyatam

International Institute of Information Technology, Hyderabad

7^{th} April 2014

Overview

- Motivation
- Problem Statement
- Related Work
- System Overview
- Approach
- Evaluation Metric
- Results
- Analysis
- Drawbacks
- Conclusion and Future Work
- References

Motivation - Page 1

- Sandhan is a mission mode project funded by TDIL, Ministry of Information Technology, Government of India.
- Its objective is to develop a mono lingual search system for tourism and health domains in ten Indian languages.
- Involves 10 Universities and 2 Government organizations.
- Visit http://tdil-dc.in/sandhan/ for a demo.

Figure 1 : Sandhan Snapshot

Motivation - Page 2

- IIIT-Hyderabad and another partner institute were responsible for crawling tourism content for all languages.
- All Language Verticals were asked to provide Seed URLs.
- General trend was that we had higher P@k values for exhaustive crawls.
- Some languages were having low P@k values despite having huge crawl.
- Upon deeper inspection we found that their Seed URL set was not diverse.

Motivation - Page 3

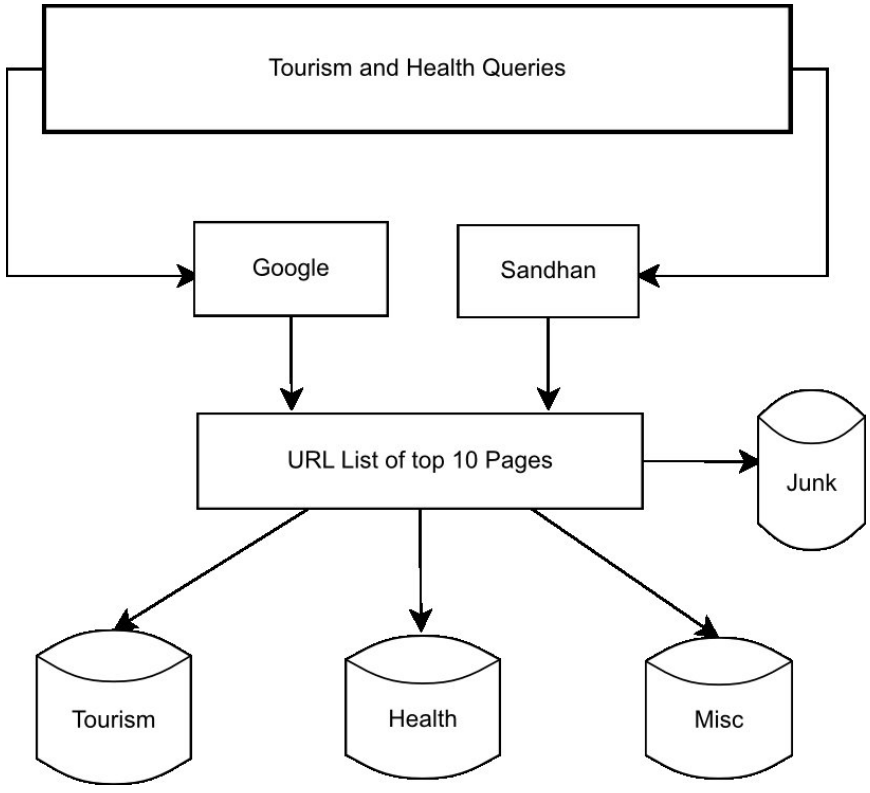

Figure 2 : Seed Set Collection

Problem Statement - Page 1

- How to maintain Seed URL diversity while building a domain specific search engine.
- People use domain specific search engines in order to get a better search experience than a generic search engine.
- In such a scenario, crawl coverage is a crucial aspect.
- A diverse seed set ensures proper crawl coverage.

Problem Statement - Page 2

Example of diverse content in tourism domain:

- Hotels
- Transportation
- Places of historic significance
- Famous cuisines
- Shopping Malls
- Weather reports
- Tourism industry
- Tourism Ministry
- Travel or Visa related info.
- ...

Related Work

- Zheng et al. [1] presented a graph based approach to select seed URLs for web crawlers.
- They employ several seed selection strategies based on PageRank, number of outlinks and website importance.

- Dmitriev et al. [2] presented a graph based approach to select seed URLs for web crawlers.
- Their seed selection algorithm takes into account popularity, trustworthiness, reliability, quality and many other parameters.

None of these works describe seed selection for domain specific search and both these works require a prior crawl or the knowledge of linkage structure.

System Overview

- We use Twitter as a Source for URLs
- We work on only English language tweets

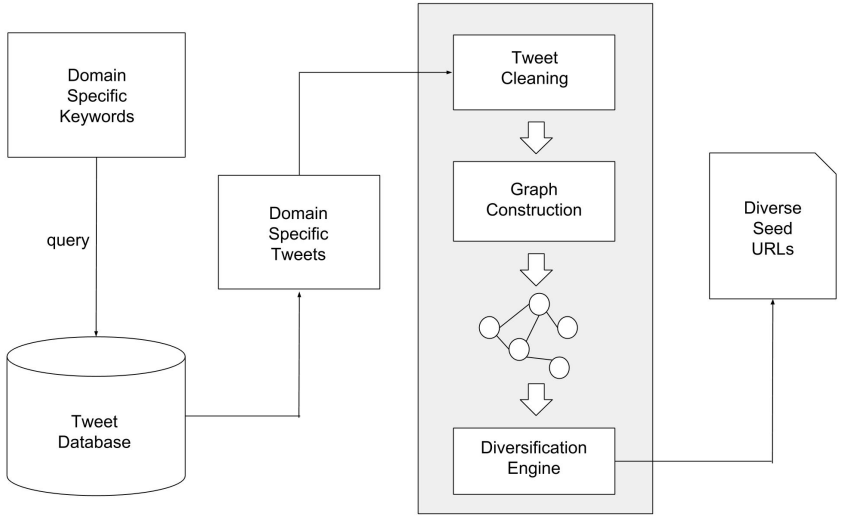

Figure 3 : System Overview

Diversification Algorithm

Algorithm 1 Diversification Algorithm

1: **Input :** *Graph* $G(V, E)$, *number of seeds* k, $k < |V|$
2: **Output :** *Diverse* k *seed URLs*
3: *Initialize Picked Nodes* $P = \{\varnothing\}$, *Eliminated Nodes* $E_l = \{\varnothing\}$, $hops = |V| - 1$
4: **while** $|P| <= k$ **do**
5: *Pick random node* n *such that* $n \in V$ *and* $n \notin P$ *and* $n \notin E_l$
6: *Add* n *to* P
7: *Add neighbours*(n, h) *to* E_l
8: **if** $|E_l \cup P| = |V|$ **then**
9: *Reinitialize* $P = \{\varnothing\}$, *Eliminated Nodes* $E_l = \{\varnothing\}$
10: $h = h - 1$
11: **end if**
12: **end while**
13: *Return* P

Figure 4 : Diversification Algorithm

Working of the Algorithm

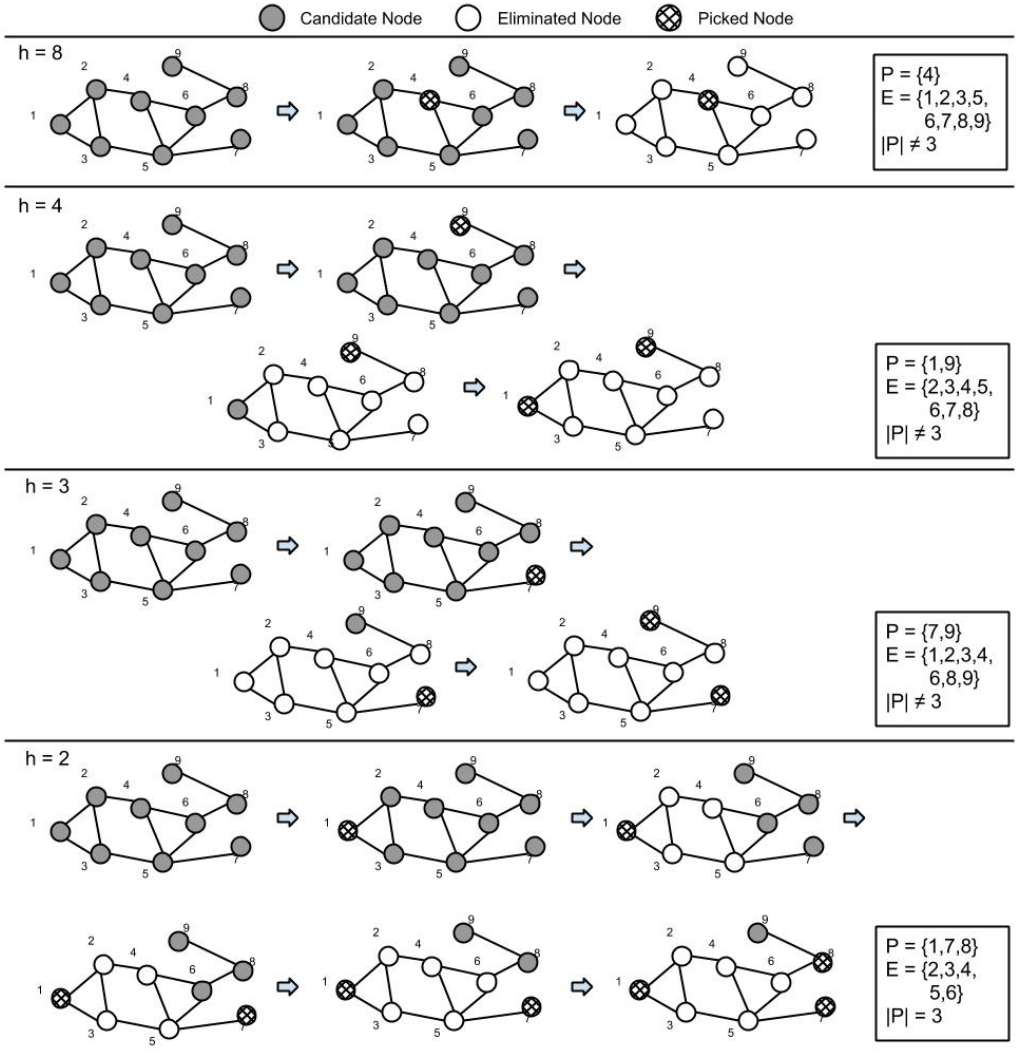

Figure 5 : Example of Graph Construction Algorithm with k=3

Graph Constrcution

The following 3 things were used to construct the graph

- Tweet Content
- Tweet URLs
- User Information

The baseline system we use is that of zero similarity - A graph with no edges.

Evaluation Metric

- The diversity of a seed URL set is judged by the diversity of the web crawl that it leads to.
- Each seed URL is manually looked at and all irrelevant URLs are removed.
- All seed sets are crawled - This is a depth 1 crawl.
- Dispersion is calculated for each seed set.

$$dispersion = \frac{\sum_{i=1}^{N}(\vec{d_i} - \vec{\mu})^2}{N} \tag{1}$$

Where $\vec{d_i}$ refers to document i represented as a bag of words vector, $\vec{\mu}$ represents the mean of all $\vec{d_i}$ $'s$ and N represents total number of documents selected.

Results

Threshold	0.15	0.20	0.25	0.3	0.35
Zero Similarity	35.0	35.0	35.0	35.0	35.0
Content	69.1	36.1	31.7	36.4	38.4
URL	34.2	39.9	42.7	33.9	39.8
User	32.0	32.0	32.0	32.0	32.0
Content + URL	49.9	47.6	34.7	59.5	42.6
User + URL	44.4	29.2	38.8	35.4	35.6
User + Content	64.2	29.4	36.6	42.0	32.7
Content + URL + User	**62.6**	**63.4**	51.4	32.6	29.6

Table 1 : Dispersion

Analysis

- Content + URL + User approach of graph construction outperforms the rest.
- Though Content, URL and user do not show clear supremacy over baseline, their combinations Content + URL, User + URL and Content + User easily beat the baseline.
- The performance of all approaches becomes randomized when the threshold reaches 0.3

Drawbacks

- Our evaluation involves manual labour and is often costly, this allowed us only depth 1 crawl.
- Our approach heavily depends on data source.
  - It cannot be extended to other languages.
  - Might not work well in less popular domains like "Nuclear Physics"
- The graph construction phase is time consuming - order of V^2 computations.

Conclusions

- We have defined the problem of seed selection for domain specific search.
- This is the 1st attempt to tap social media for solving the problem of seed selection.
- This work gives special treatment to the problem of capturing diversity in domain-specic crawling.
- This field has lot of promise and will become essential in future.

Future Work - Page 1

- Instead of picking randomly pick the most relevant: provided we have a metric to measure relevance of URLs to the domain.
- We would want to try this approach on a different dataset like Facebook.
- We would like to study diversity measure of a seed set at different time slices of the crawl.
- We would like to compare our seed URLs with existing sets of domain specic URLs like the ones present in Open Directory Project [1].

[1] http://www.dmoz.org/

Future Work - Page 2

- We also aim to compare against manual seed selection using crowdsourcing platforms like *Amazon Mechanical Turk* [2] and *CrowdFlower* [3].
- Use sub-topic or sub-domain structure to evauate the diversity of the crawl.

[2] http://www.mturk.com
[3] http://crowdflower.com/

References

[1] S. Zheng, P. Dmitriev, and C. Giles. Graph based crawler seed selection. In Proceedings of the 18th international conference on World wide web, pages 1089–1090. ACM, 2009.

[2] P. Dmitriev. Host-based seed selection algorithm for web crawlers, 2008. US Patent App. 12/259,164.