Domain Specific Search in Indian Languages

P Nikhil Priyatam
Srikanth Reddy
Vasudeva Varma

International Institute of Information Technology, Hyderabad, India

November 2, 2012

Outline

- Overview of Sandhan
- Introduction
- Previous Approaches
- Our Approach
- Results
- Conclusions and Future Work

Overview of Sandhan

Goal: To offer cross lingual search in the domains Tourism and Health across 10 different Indian languages.
Link: http://tdil-dc.in/sandhan/

Domain specific search engines

A domain specific search engine indexes pages belonging to a specific domain and discards others. There are two ways to do this:

- Using an unfocused crawl and a Domain Identifier: In this strategy we crawl all pages and classify them as relevant or irrelevant using a domain classifier (a Web page classification algorithm). The relevant pages are indexed and irrelevant ones are discarded.
- Using Focused crawler: A Focused crawler on the other hand identifies URLs that are great access points to many relevant pages and fetches only those pages which are relevant.

Types of Focused crawlers

Focused crawlers can be categorized into two types:

- Close domain: A.K.A topic specific crawlers. These crawlers fetch pages belonging to a specific topic like bycycling, specific disease etc.
- Open domain: A.K.A domain specific crawlers. These crawlers fetch pages belonging to a particular domain like tourism, health, sports etc.

Previous Approaches

- Chakrabarti et al built a classifier guided focused crawler, using a simple crawling strategy: If a page is relevant so might be its outlinks (hyperlinks). They reported a precision of 0.4, but on a limited set of 10,000 pages.
- Medelyan et al used similar strategy for the medical domain, however they report the precision of their classifier rather than the performance of the focused crawler.
- Mukhopadhyay and stamatakis et al use ontology based web crawling.
- Aggarwal et al provided a framework for intelligent crawling where the crawler gradually learns the linkage structure as it progresses
- Menczer et al gave the evaluation metrics for focused crawlers.

Drawbacks of previous approaches

- Most of the previous approaches report precision of their focused crawlers but none of them talks about recall or crawl coverage.
- A good crawl coverage is essential to offer domain specific search.
- Most of these approaches work for close domains but do not work on open domains like tourism or health [Fesenmeier et al].
- No prior work has been done in building focused crawlers for Indian languages.

Challenges in building Indian language focused crawlers

The combination of an Indian language and an open domain poses the following challenges:

- Scarcity of content
- Proprietary Fonts
- Other language Content
- Language identification
- Lack of training data.

Pingali et al proposed a working approach to overcome some of these challenges, but they do not address issues related to domain specific search.

Our Approach

- In this work we build a classifier guided focused crawler for Hindi tourism and health pages.
- We decided to use a Naive bayes classifier which classifies a page into tourism,health and miscellaneous.
- For the classifier to work well on the real life data we need to model the actual distribution.
- Therefore the trick is in collecting data.

Data Collection Step 1: Tourism

We first decide what information should a tourism or health document contain.

- Information about historical places
- Tourist attractions
- Travel Guide which includes cost of trip,best time to visit, transport facilities, nearby hotels, accommodation facilities, and nearby places of interest of a particular tourist spot.
- Food and other popular cuisines.
- On line services (if any) provided by the respective tourist spots.
- Latest news about a particular place which might include weather reports and any other news alerts.

Data Collection Step 1: Health

- Complete information about any disease: Symptoms, precautions, cure and other related information.
- Information about medicines, their ingredients, possible side effects, availability and the type of medicine (ayurvedic, allopathy, homeopathy, etc.)
- Information about clinics or hospitals, doctors etc.
- Latest research news about some ailments.
- Information related to nutrition, diet, personal hygiene, etc.

Data Collection Step 2: Query Collection

Around 30 people from different places (states) in India were asked to provide the following:

- 3 regional queries and their translations in English and Hindi.
- 3 non-regional queries and their translations in English and Hindi.
- 3 health queries and their translations in English and Hindi.

In this way we were able to collect 110 tourism queries and 60 health queries.

Firing Queries onto Search engine

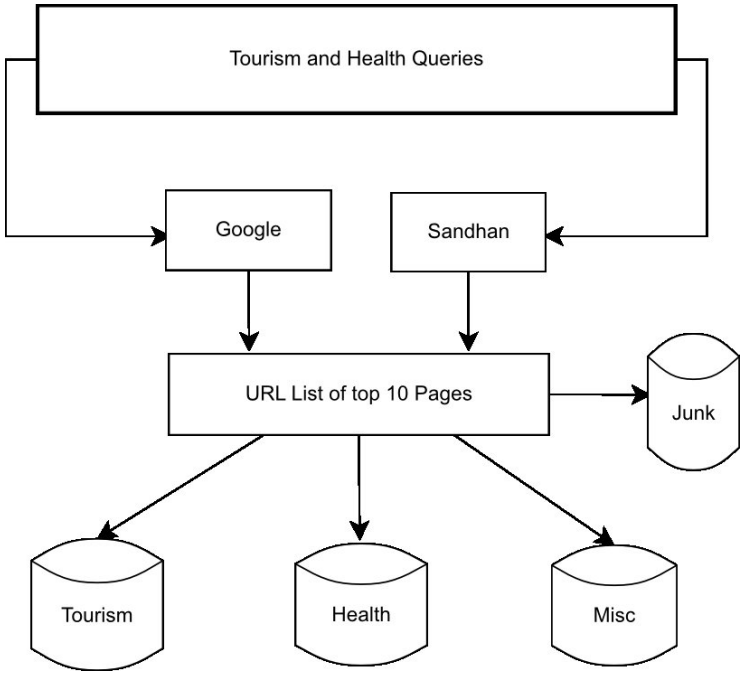

Table: Statistics of data collected

Domain	No of Web Pages
Tourism	885
Health	978
Miscellaneous	1354

Web page classification

Table: Confusion Matrix of Naive Bayes classifier

Tourism	Health	Miscellaneous	Precision	Recall	F-Score	
502	48	78	0.85	0.8	0.83	Tourism
6	500	45	0.75	0.91	0.82	Health
80	121	707	0.85	0.78	0.81	Miscellaneous

Focused crawler

Defn: A web crawler that selectively seeks out pages relevant to a predefined domain.

- Using the WPC mentioned above we decide the relevance of a page.
- We use a simple crawling strategy i.e: if a page is relevant so might be its outlinks.

Intuition behind our approach: We are assuming 2 properties about the nature of the web:

- Linkage locality: Web pages on a given topic are more likely to link to those on the same topic.
- Sibling locality: If a web page points to certain web pages on a given topic, then it is likely to point to other pages on the same topic.

Evaluation Metrics

We evaluate the quality of the crawl using 3 metrics: precision, recall and harvest ratio.

$$Precision = \frac{\#\ of\ relevant\ pages}{Total\ \#\ of\ pages\ fetched} \tag{1}$$

$$Recall = \frac{\#\ of\ relevant\ pages\ fetched\ by\ focused\ crawler}{Total\ \#\ of\ relevant\ pages} \tag{2}$$

$$Recall = \frac{\#\ of\ relevant\ pages\ fetched\ by\ focused\ crawler}{\#\ of\ relevant\ pages\ fetched\ by\ unfocused\ crawler} \tag{3}$$

Harvest ratio measures the average number of relevant pages retrieved over different time slices of the crawl.

Results

The URLs of web pages collected for building the classifier were used as seed URLs to crawl till a depth of 3.

Table: Quality of crawl

Crawler	Precision	Recall
Unfocused crawler tourism	0.105	
Unfocused crawler health	0.106	
Focused tourism	0.4	0.74
Focused health	0.36	0.58

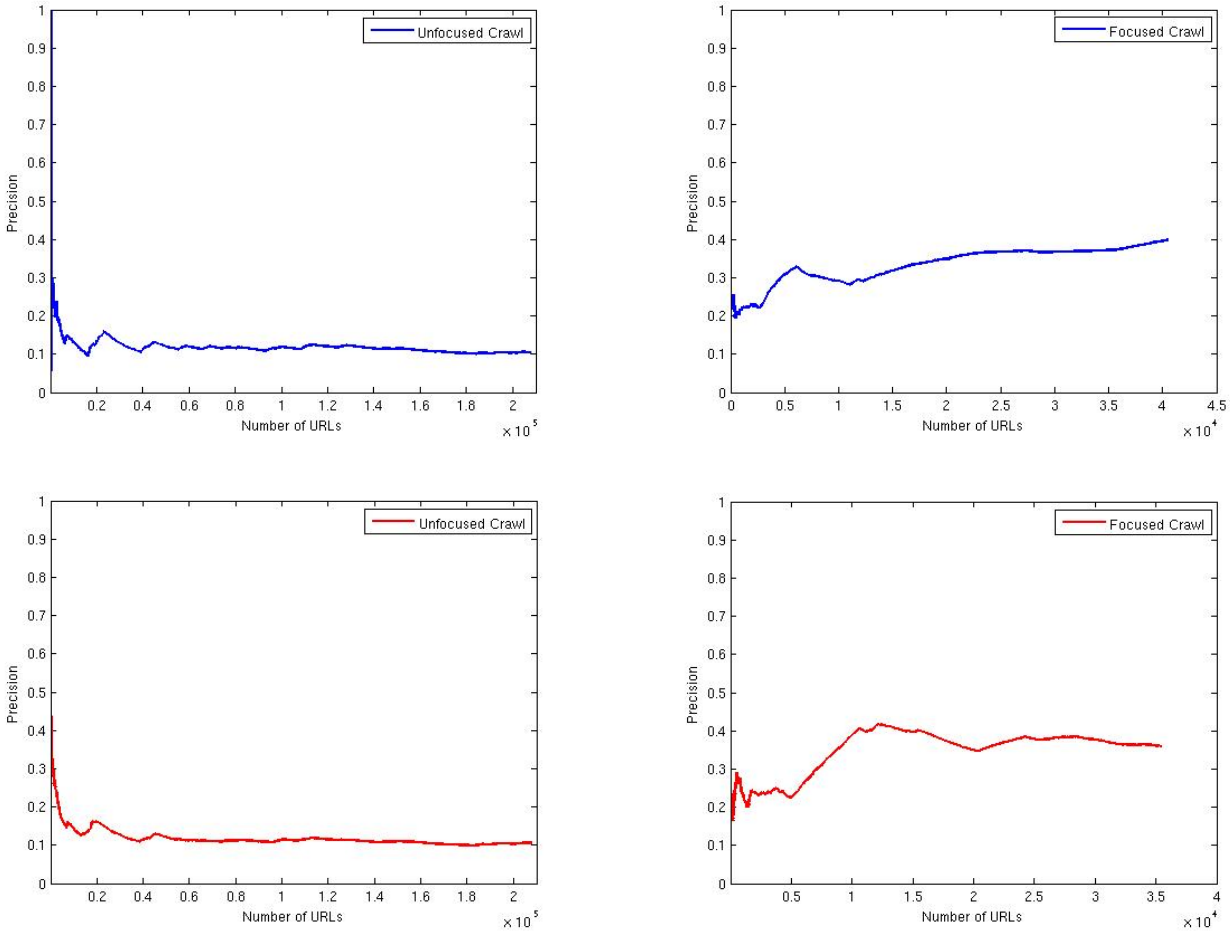

Figure: Performance comparison of unfocused and focused crawler in tourism and health domains. Blue and red color indicate tourism and health domains respectively.

Observations

Table: Language Statistics of Focused and Unfocused crawler

	Unfocused crawler	Focused Crawler-tourism	Focused Crawler-health
Hindi	94995	31182	21298
English	83234	6557	9229
Gujarati	386	15	47
Punjabi	17	0	3
Bengali	275	3	21
Assamese	82	10	10
Marathi	921	220	152
Telugu	1543	12	59
Tamil	278	3	37
Oriya	1	0	0
Unidentified	26298	2530	4346

Focused crawler fetches less number of other language pages.

Conclusions

- In this work we build a classifier guided focused crawler to fetch Hindi tourism and Health pages.
- Our data collection method ensures that training and testing datasets come from the same distribution.
- The focused crawler for health achieves a huge recall of 0.58, which means that it was able to fetch nearly 60 % of all relevant pages with a crawl size of 16 % of exhaustive crawl.
- Similarly, The focused crawler for tourism achieves a huge recall of 0.74, which means that it was able to fetch nearly 75 % of all relevant pages with a crawl size of 20 % of exhaustive crawl.
- In both cases we save nearly 80 % of the resources.

Future Work

- Elimination of other language pages from the crawl using URL based language identification.
- Moving from all-or-none strategy to selective pick strategy.
- Explore different feature spaces for building domain identifier
- Using other crawling strategies.

ANY QUESTIONS / QUERIES ?

THANK YOU.