Developing Focused Crawlers for Genre Specific Search Engines

Nikhil Priyatam

Thesis Advisor: Prof. Vasudeva Varma

IIIT Hyderabad

July 7, 2014
Examples of Genre Specific Search Engines

- MedlinePlus
- Naukri.com
- DirectIndustry
- RecipeBridge
- MobileWalla
Outline

- Introduction
- Focused Crawlers for Sandhan
- URL Based Genre Identification
- Seed Selection for Genre Specific Search
- Measuring Diversity of a Genre Specific Crawl
Figure: Overview of a Web Search Engine

1Source: http://www.ibm.com/developerworks/library/wa-lucene2/
Figure: Difference between a Normal Crawler and a Focused Crawler
Challenges

- How to decide the relevance of a web page?
- How to achieve a high recall? A severe problem for Indian language genre specific search engines.
- How to select seeds for genre specific search?
- Which crawling strategy to use?
Outline

- Introduction
- Focused Crawlers for Sandhan
- URL Based genre Identification
- Seed selection for genre specific search
- Measuring Diversity of a Genre Specific Crawl
There are two ways to gather content for Sandhan:

- Using a crawler and a genre identifier or
- Building a focused crawler.
Drawbacks of previous approaches

None of the previous approaches can be directly used for Sandhan due to the following reasons:

- 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 the existing approaches work for close domains but do not work on open domains like tourism or health \(^2\).
- No prior work has been done in building focused crawlers for Indian languages.

\(^2\)D. R. Fesenmaier. *Domain Specific search engines*
The combination of an Indian language and an open domain poses the following challenges:

- Proprietary Fonts
- Other language Content
- Language identification
- Scarcity of content
- 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.
In this work we build a classifier guided focused crawler for Hindi tourism and health pages.

We decided to use a Naïve bayes classifier which classifies a page into tourism, health and general.

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.
Define the domains: tourism and health

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

<table>
<thead>
<tr>
<th>Domain</th>
<th>No of Web Pages</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tourism</td>
<td>885</td>
</tr>
<tr>
<td>Health</td>
<td>978</td>
</tr>
<tr>
<td>Miscellaneous</td>
<td>1354</td>
</tr>
</tbody>
</table>
Web page classification

<table>
<thead>
<tr>
<th></th>
<th>Tourism</th>
<th>Health</th>
<th>Miscellaneous</th>
<th>Precision</th>
<th>Recall</th>
<th>F-Score</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Tourism</strong></td>
<td>502</td>
<td>48</td>
<td>78</td>
<td>0.85</td>
<td>0.8</td>
<td>0.83</td>
</tr>
<tr>
<td><strong>Health</strong></td>
<td>6</td>
<td>500</td>
<td>45</td>
<td>0.75</td>
<td>0.91</td>
<td>0.82</td>
</tr>
<tr>
<td><strong>Miscellaneous</strong></td>
<td>80</td>
<td>121</td>
<td>707</td>
<td>0.85</td>
<td>0.78</td>
<td>0.81</td>
</tr>
</tbody>
</table>

Table: Confusion Matrix of Naïve Bayes classifier
We use a simple crawling strategy i.e: if a page is relevant so might be its outlinks.

Relevance of a page is judged using the WPC mentioned before.

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.

\[
Precision = \frac{\text{\# of relevant pages}}{\text{Total \# of pages fetched}}
\]  

(1)

\[
Recall = \frac{\text{\# of relevant pages fetched by focused crawler}}{\text{Total \# of relevant pages}}
\]  

(2)

\[
Recall = \frac{\text{\# of relevant pages fetched by focused crawler}}{\text{\# of relevant pages fetched by unfocused crawler}}
\]  

(3)

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

<table>
<thead>
<tr>
<th>Crawler</th>
<th>Precision</th>
<th>Recall</th>
</tr>
</thead>
<tbody>
<tr>
<td>Unfocused crawler tourism</td>
<td>0.105</td>
<td>1</td>
</tr>
<tr>
<td>Unfocused crawler health</td>
<td>0.106</td>
<td>1</td>
</tr>
<tr>
<td>Focused tourism</td>
<td>0.4</td>
<td>0.74</td>
</tr>
<tr>
<td>Focused health</td>
<td>0.36</td>
<td>0.58</td>
</tr>
</tbody>
</table>
Figure: Performance comparison of unfocused and focused crawler in tourism and health domains. Blue and red colours indicate tourism and health domains respectively.
Outline

- Introduction
- Focused Crawlers for Sandhan
- **URL Based genre Identification**
- Seed selection for genre specific search
- Measuring Diversity of a Genre Specific Crawl
URL based methods have several advantages and they should be employed when:

- Classification speed must be high.
- Content filtering is needed before an objectionable or advertisement page is downloaded.
- Page’s content is hidden in images or non-standard encodings.
- Annotation needs to be performed on hyperlinks in a personalized web browser, without fetching the target page.
- Focused crawler wants to infer the topic of a target page before devoting bandwidth to download it.
- Language of the page needs to be identified.
No single work reported for Indian languages

Most of the existing approaches require huge amount of resources such as heavy training data, already existing corpus or web graph information collected from a search engine.

Many works like [1] report their results on toy datasets and hence are not scalable to be used in search engines like Sandhan.

In this work we build a URL based genre identification system on a huge dataset using minimal resources and training.

---

1. What's in a URL? Genre Classification from URLs
2. Topical Host Reputation for Lightweight URL Classification
Previous Approaches

- Baykan et al.\(^1\) uses a combination of character N-grams and SVM to classify URLs into one of 15 predefined categories.
- Abramson et al.\(^2\) train a classifier on selectively picked character n-grams (4 to 8) by eliminating n-grams which are redundant in nature.
- Kan et al.\(^3\) train a classifier on an exhaustive set of features which include character n-grams, sequential n-grams, URL components, URL tokens length and orthographic features like number of queries and number of numerics in the URL.
- Hernndez et al.\(^4\) use a clustering based approach to identify URL category.

---

1. Purely URL-Based Topic Classification
2. What's in a URL? Genre Classification from URLs
3. Fast Webpage Classification Using URL Features
4. A Statistical Approach to URL-Based Web Page Clustering
Dataset and System Architecture

- **Dataset**: 94995 Hindi web pages tagged as tourism/health/misc by a WPC (described above)
- **Features**:
  - Words
  - n-grams (4 to 8)
  - All grams
We use 3 approaches for URL based genre identification

- List Based
  - Manually prepared list
  - via an external corpus
  - via a retrieval system

- Naïve Bayes

- Incremental Naïve Bayes (Outperforms the other two)
## Results

### Table: Incremental Naïve Bayes Approach

<table>
<thead>
<tr>
<th>Genre</th>
<th>Tourism</th>
<th>Health</th>
</tr>
</thead>
<tbody>
<tr>
<td>Feature</td>
<td>P</td>
<td>R</td>
</tr>
<tr>
<td>Words</td>
<td>0.849</td>
<td>0.834</td>
</tr>
<tr>
<td>4 grams</td>
<td>0.849</td>
<td>0.824</td>
</tr>
<tr>
<td>5 grams</td>
<td>0.858</td>
<td>0.862</td>
</tr>
<tr>
<td>6 grams</td>
<td>0.860</td>
<td>0.873</td>
</tr>
<tr>
<td>7 grams</td>
<td>0.860</td>
<td>0.880</td>
</tr>
<tr>
<td>8 grams</td>
<td>0.858</td>
<td>0.886</td>
</tr>
<tr>
<td>All grams</td>
<td>0.856</td>
<td>0.869</td>
</tr>
</tbody>
</table>
Comparison

<table>
<thead>
<tr>
<th>Approach</th>
<th>Tourism</th>
<th>Health</th>
</tr>
</thead>
<tbody>
<tr>
<td>Incremental Naïve Bayes</td>
<td>0.858</td>
<td>0.873</td>
</tr>
<tr>
<td>Baykan et al.</td>
<td>0.81</td>
<td>0.37</td>
</tr>
<tr>
<td>Abramson et al.</td>
<td>0.82</td>
<td>0.39</td>
</tr>
<tr>
<td>Kan et al.</td>
<td>0.77</td>
<td>0.30</td>
</tr>
</tbody>
</table>

Table: Comparison of Incremental Naïve Bayes Approach and Earlier Approaches proposed in Literature Survey

1Purely URL-Based Topic Classification
2Whats in a URL? Genre Classification from URLs
3Fast Webpage Classification Using URL Features
Outline

- Introduction
- Focused Crawlers for Sandhan
- URL Based genre Identification
- Seed selection for genre specific search
- Measuring Diversity of a Genre Specific Crawl
Coverage and diversity of crawl are two of the pivotal aspects of a genre specific search engine mostly governed by the initial set of seed URLs.

Generally, seed URLs are collected manually.

Following factors influence the quality of seed set:

- Relevance
- Diversity of the seed set
- Seed set size

This task requires domain expertise and manual effort. No automated system exists for this purpose.

In this work, we present an approach to automate the process of seed URLs collection for domain specific search with a special focus on “diversity”. For this we use twitter data.
Why Twitter?

- About 25% tweets contain URLs
- Users post and exchange information about a variety of trending topics like entertainment, politics, tourism, etc.
- Twitter has millions of users coming from different social, geographical and cultural backgrounds which ensures a diverse audience.
- Twitter provides to the users the option of following other users.
- Content similarity/overlap between tweets of different users gives us the diversity of user opinions.
- The trail of tweets over time can be used to ensure temporal diversity.
- Huge number of new URLs are posted everyday. Following these URLs would lead to a fresh and updated crawl.

http://techcrunch.com/2010/09/14/twitter-event/
**Genre specific keywords:** Manually fed into the system

**Twitter Search:** We have searched Nov 2009 tweets

**Tweet Validation:** Filters tweets containing valid URLs

**Graph construction:** Constructs an undirected unweighted graph from the validated tweets

**Diversification Engine:** Returns k diverse seed URLs
Algorithm 1 Diverse Seed Selection Algorithm

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

Figure: Example of Algorithm 1, Number of Seeds $k=3$
Graph Construction

- Zero Similarity (Baseline): No two vertices are connected in the graph.
- Content Similarity: Two URLs are connected if the tweets that contain these URLs have content overlap above a threshold.
- URL N-Grams Similarity: Two URLs are connected if the URL n-grams have overlap above a threshold.
- User Similarity: Two users are considered similar if at least one of them has retweeted the other’s tweet.
The diversity of a seed URL set is judged by the diversity of the web crawl that it leads to.

We measure the variance across the crawled set of documents to judge its diversity, called Dispersion.

Dispersion is average squared distance of all documents from the mean

$$dispersion = \frac{\sum_{i=1}^{N} (\vec{d}_i - \vec{\mu})^2}{N}$$  \hspace{1cm} (4)

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.
<table>
<thead>
<tr>
<th>Threshold</th>
<th>0.15</th>
<th>0.20</th>
<th>0.25</th>
<th>0.3</th>
<th>0.35</th>
</tr>
</thead>
<tbody>
<tr>
<td>Zero Similarity</td>
<td>35.0</td>
<td>35.0</td>
<td>35.0</td>
<td>35.0</td>
<td>35.0</td>
</tr>
<tr>
<td>Content</td>
<td>69.1</td>
<td>36.1</td>
<td>31.7</td>
<td>36.4</td>
<td>38.4</td>
</tr>
<tr>
<td>URL</td>
<td>34.2</td>
<td>39.9</td>
<td>42.7</td>
<td>33.9</td>
<td>39.8</td>
</tr>
<tr>
<td>User</td>
<td>32.0</td>
<td>32.0</td>
<td>32.0</td>
<td>32.0</td>
<td>32.0</td>
</tr>
<tr>
<td>Content + URL</td>
<td>49.9</td>
<td>47.6</td>
<td>34.7</td>
<td>59.5</td>
<td>42.6</td>
</tr>
<tr>
<td>User + URL</td>
<td>44.4</td>
<td>29.2</td>
<td>38.8</td>
<td>35.4</td>
<td>35.6</td>
</tr>
<tr>
<td>User + Content</td>
<td>64.2</td>
<td>29.4</td>
<td>36.6</td>
<td>42.0</td>
<td>32.7</td>
</tr>
<tr>
<td>Content + URL + User</td>
<td><strong>62.6</strong></td>
<td><strong>63.4</strong></td>
<td>51.4</td>
<td>32.6</td>
<td>29.6</td>
</tr>
</tbody>
</table>

**Table:** Dispersion Values for Different Approaches
<table>
<thead>
<tr>
<th>Seed Selection Source</th>
<th>Dispersion</th>
</tr>
</thead>
<tbody>
<tr>
<td>Twitter (using Algorithm 1)</td>
<td>63.4</td>
</tr>
<tr>
<td>Web (using Manual Collection)</td>
<td><strong>84.82</strong></td>
</tr>
<tr>
<td>ODP</td>
<td>25.83</td>
</tr>
<tr>
<td>Wikipedia</td>
<td>30.62</td>
</tr>
</tbody>
</table>

**Table:** Comparison of Seed Collection with other Types of Seeds
Outline

- Introduction
- Focused Crawlers for Sandhan
- URL Based genre Identification
- Seed selection for genre specific search
- Measuring Diversity of a Genre Specific Crawl
While measuring diversity of Seed Sets we used only one metric: dispersion. Here we propose 3 new metrics to measure the diversity of a genre-specific crawl.

These metrics can be used to evaluate focused crawlers and seed sets w.r.t diversity.

- Semantic Distance
- Average Cosine Similarity
- KL-divergence using Topic Models
Semantic Distance

\[ SD(d_x, d_y) = \sum_{i=1}^{k} \sum_{j=1}^{k} \text{WordNet Distance}(w_{xi}, w_{yj})^4 \]  \hspace{1cm} (5)

where \(SD(d_x, d_y)\) represents the semantic distance between documents \(x\) and \(y\) respectively and \(w_{xi}, w_{yj}\) represents the \(i^{th}\) word of document \(x\) and the \(j^{th}\) word of document \(y\) respectively.

\[
D1 \ \text{Score} = \frac{\sum_{x=1}^{N} \sum_{y=1}^{N} SD(d_x, d_y)}{N^2} \hspace{1cm} (6)
\]

\(^4\)http://rednoise.org/rita/wordnet/documentation/riwordnet
Average Cosine Similarity

\[ \text{ACS} = \frac{\sum_{i=1}^{k} \sum_{j=1, j \neq i}^{k} \text{Cosine Similarity}(d_i, d_j)}{\text{number of unique unordered document pairs}} \]  \hspace{1cm} (7)

\[ D3 \text{ Score} = \frac{1}{\text{ACS}} \]  \hspace{1cm} (8)
KL-Divergence using Topic Models

\[ D4 \text{ Score} = \sum_{i=1}^{k} \sum_{j=1, j \neq i}^{k} \text{KLDivergence}(t_i, t_j) \]  \hspace{1cm} (9)

where \( t_i \) and \( t_j \) represent topic \( i \) and topic \( j \), and

\[ \text{KLDivergence}(t_i, t_j) = \sum_{v=1}^{|V|} \ln \left( \frac{t_i(v)}{t_j(v)} \right) t_i(v) \]  \hspace{1cm} (10)

where \( t_i(v) \) and \( t_j(v) \) represent the probabilities of word \( v \) in topics \( i \) and \( j \) respectively and \( |V| \) represents vocabulary size. Hence a web crawl covering varied topics will have a higher diversity score than the crawl containing similar topics.
Feature Space

For dispersion (equation 4) and average cosine similarity (equation 7) based metric we use the following two feature spaces:
1. Bag of Words (TF-IDF)
2. Context Vectors
The algorithm to compute context vectors of a document is given below:

**Algorithm 2 Context Vector Algorithm**

1: **Input**: Document $d$, number of word vectors $n$, window length $k$
2: **Output**: Context Vector of document $d$
3: Calculate $tf-idf$ scores for all words in $d$
4: Pick $n$ words with highest $tf-idf$ values call it $W$
5: for each word $w_i$ in $W$ do
6: 
7: $\vec{w}_i =$ Vector containing $tf-idf$ values of all words in $w_i^k$
8: end for
9: $CV = \text{Centroid}(\vec{w}_1, \vec{w}_2, \ldots, \vec{w}_n)$
10: Return $CV$
How to evaluate Metrics

- We evaluate the usefulness of a metric by its ability to distinguish between More diverse and Less diverse web crawl.
- We experiment on 3 domains: Tourism, Health and Sports.

Figure: Picking URLs from ODP Hierarchy
### Results

<table>
<thead>
<tr>
<th>Feature Space</th>
<th>Bag of Words</th>
<th>Context Vectors</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>MD</td>
<td>LD</td>
</tr>
<tr>
<td>Tourism</td>
<td>53.19</td>
<td>31.35</td>
</tr>
<tr>
<td>Health</td>
<td>49.54</td>
<td>18.61</td>
</tr>
<tr>
<td>Sports</td>
<td>37.60</td>
<td>14.07</td>
</tr>
</tbody>
</table>

**Table: Average Semantic Distance based Metric**

<table>
<thead>
<tr>
<th>Feature Space</th>
<th>Bag of Words</th>
<th>Context Vectors</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>MD</td>
<td>LD</td>
</tr>
<tr>
<td>Tourism</td>
<td>0.788</td>
<td>0.779</td>
</tr>
<tr>
<td>Health</td>
<td>0.759</td>
<td>0.721</td>
</tr>
<tr>
<td>Sports</td>
<td>0.764</td>
<td>0.769</td>
</tr>
</tbody>
</table>

**Table: Topic Modeling based Metric**

<table>
<thead>
<tr>
<th>Feature Space</th>
<th>Bag of Words</th>
<th>Context Vectors</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>MD</td>
<td>LD</td>
</tr>
<tr>
<td>Tourism</td>
<td>118.31</td>
<td>109.42</td>
</tr>
<tr>
<td>Health</td>
<td>169.58</td>
<td>110.56</td>
</tr>
<tr>
<td>Sports</td>
<td>118.55</td>
<td>100.48</td>
</tr>
</tbody>
</table>

**Table: Similarity based Metric**

<table>
<thead>
<tr>
<th>Feature Space</th>
<th>Bag of Words</th>
<th>Context Vectors</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>MD</td>
<td>LD</td>
</tr>
<tr>
<td>Tourism</td>
<td>118.55</td>
<td>100.48</td>
</tr>
<tr>
<td>Health</td>
<td>169.58</td>
<td>110.56</td>
</tr>
<tr>
<td>Sports</td>
<td>118.31</td>
<td>109.42</td>
</tr>
</tbody>
</table>

**Table: Dispersion based Metric**
Comparison with other Diversity Measures

<table>
<thead>
<tr>
<th>Approach</th>
<th>Tourism</th>
<th>Health</th>
<th>Sports</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>MD</td>
<td>LD</td>
<td>Ratio</td>
</tr>
<tr>
<td>Similarity Based Metric</td>
<td>83.33</td>
<td>46.26</td>
<td>1.77</td>
</tr>
<tr>
<td>Refined Diversity JaccardIndex</td>
<td>33.33</td>
<td>27.02</td>
<td>1.23</td>
</tr>
<tr>
<td>Sampled Diversity JaccardIndex</td>
<td>32.25</td>
<td>29.41</td>
<td>1.09</td>
</tr>
<tr>
<td>Simpsons Inverse Diversity</td>
<td>3.717</td>
<td>3.533</td>
<td>1.05</td>
</tr>
</tbody>
</table>

**Table:** Comparison of Metrics to Measure Diversity

---

5 Efficient Jaccard-Based Diversity Analysis of Large Document Collections

6 Efficient Jaccard-Based Diversity Analysis of Large Document Collections

7 Measurement of diversity
Built Focused Crawler for Sandhan

Established the use of incremental methods in the context of Focused Crawling.

We were the first ones to automate the process of Seed selection for a genre specific search.

Proposed various metrics to measure the diversity of a genre specific web crawl.
Questions?