Don't Use a Lot When Little Will Do : Genre Identi cation Using URLs

by

Nikhil Priyatam, Srinivasan Iyengar, Krish Perumal, Vasudeva Varma

in

14th International Conference on Intelligent Text Processing and Computational Linguistics (CICLing 2013)

University of the Aegean, Samos, Greece

Report No: IIIT/TR/2013/-1

Centre for Search and Information Extraction Lab
International Institute of Information Technology
Hyderabad - 500 032, INDIA
March 2013

Don't Use a Lot When Little Will Do : Genre Identification Using URLs

Pattisapu Nikhil Priyatam[1], *Srinivasan Iyengar*[2], *Krish Perumal*[1] *and Vasudeva Varma*[1]
nikhil.priyatam@research.iiit.ac.in,
venkatachary.srinivasaniyengar@tcs.com,
{krish.perumal@research.,vv@}iiit.ac.in

(1) Search and Information Extraction Lab, IIIT-Hyderabad, India
(2) Tata Research Development and Design Centre, TCS-Pune, India

Abstract. The ever increasing data on world wide web calls for the use of vertical search engines. *Sandhan* is one such search engine which offers search in tourism and health genres in more than 10 different Indian languages. In this work we build a URL based genre identification module for *Sandhan*. A direct impact of this work is on building focused crawlers to gather Indian language content. We conduct experiments on tourism and health web pages in **Hindi** language. We experiment with three approaches - list based, naive Bayes and incremental naive Bayes. We evaluate our approaches against another web page classification algorithm built on the parsed text of manually labeled web pages. We find that incremental naive Bayes approach outperforms the other two. While doing our experiments we work with different features like words, n-grams and all grams. Using n-gram features we achieve classification accuracies of 0.858 and 0.873 for tourism and health genres respectively.

Keywords: Genre Identification, Focused Crawlers, Web Page Classification

1 Introduction

The online content and consumption of Indian language web pages is increasing at a rapid pace[1]. This will only increase further as broadband and mobile-based Internet user base is expanding [1]. Thus, information retrieval problems such as Focused Crawling, Web Page Classification and Web Page Ranking need to be addressed for Indian languages. Even though these problems are addressed to a sufficient degree of satisfaction for English, a lot needs to be done when it comes to Indian languages. Efforts such as IndoWordNet [2] are trying to build excellent quality language resources. Still there is a clear scarcity of language resources for Indian languages. Many Indian language web pages use proprietary non-unicode fonts or non-standard character encodings. This is owing to

[1] Vernacular Content Market in India, Dec 2008 : http://www.iamai.in/rsh_pay.aspx?rid=XsZAUn657BU=

the fact that many web pages existed even before the unicode became a standard and thus have continued to use non-standard encodings. As building separate transcoders[2] for different non-standard fonts is effort intensive, most of the content of these web pages is rendered useless for many information retrieval tasks. In Information Retrieval, Web Page Classification (WPC) has many significant applications and is especially pivotal in building genre specific search engines. *Sandhan*[3] is one such search engine which offers search for tourism and health genres across 10 different Indian languages. In this work, we describe a light-weight, genre specific web page classification algorithm for Hindi. Our approach looks at clues in the URL of a web page to decide its genre. We test our approach on two genres - tourism and health. As a byproduct, we have other added advantages that come with this approach. These include classification speed and economy of operation in terms of bandwidth, processing and storage.

The flow of this paper is as follows. We discuss relevant literature on the use of URLs for different IR activities in section 2. In section 3 we describe our experimental dataset. In section 4 we describe our system architecture. We propose different approaches for WPC using URLs in section 5. Section 6 describes our experimental setup. In section 7 we present the metrics used for evaluation. In section 8 we report our results when evaluated against a WPC algorithm [3]. We conclude our discussion with future directions in section 9.

2 Literature Survey

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.

Baykan et al. [4] describe an approach to classify web pages into 15 different topics such as sports, news, adult, shopping, etc. for English language pages. Their feature list consists of topic specific words and their n-grams. They train separate binary classifiers (SVM) for each topic. Although they use words from the first two levels of the ODP hierarchy for a particular topic, some manual post processing was done to remove non-topic-specific words such as "online" and "games".

Baykan et al. [5] describe a method to identify the language of the web page by analyzing the URL. They apply a variety of machine learning

[2] A transcoder converts non-unicode text into unicode

[3] http://www.clia.iitb.ac.in:8080/sandhan-prsg/locale.jsp?en

algorithms to the language identification task and evaluate their performance in extensive experiments for five languages: English, French, German, Spanish and Italian.

Kan et al. [6] introduce features such as position, length and sequence of tokens in a URL for the task of web page classification.

Shih et al. [7] propose new features and algorithms for automating web page classification tasks such as content recommendation and ad blocking. Apart from analyzing the URL tokens, they study the placement of these links in the referring page i.e. its HTML tree structure. They develop a machine learning model and algorithm using such features.

Kan et al. [8] present a couple of techniques to segment and expand URLs into tokens for performing effective web page categorization. Initial baseline segmentation is done using non-alphanumeric characters, uppercase-lowercase transitions and digits. Further segmentation of individual chunks is done using the information content of every possible partition. The probabilities required for this information theory based method were estimated by observing document frequency of tokens over several million web pages. The chunks are further split and expanded based on previously seen web page titles using a weighted non-deterministic finite-state transducer.

Hernndez et al. [9] mention an unsupervised method to build URL patterns for web page clustering.

Anastácio et al. [10] talk about categorizing documents according to their implicit locational relevance. One of the features they use is URL n-grams. They vary n between 4 and 8 and assign weights according to the TF-IDF scheme.

Ma et al. [11] use lexical and host-based features of URLs to identify spam sites. Further they use online learning algorithms to classify malicious URLs with high accuracy. Toyoda et al. [12] further classify them into different spam topics. Kołcz et al. [13] use web-graph information along with the basic URL-based approach and show improved results for classifying webpages.

Abramson et al. [14] give an exhaustive list of features that can be extracted from a URL, which include syntactic style features, semantic style features, part-of-speech (POS) tags, punctuations and special characters. They build classifiers (Naive Bayes and SVM) using *Santini* and *Syr7* datasets.

Charu et al. [15] provide a framework for intelligent crawling where the crawler gradually learns the linkage structure as it progresses. Their crawler incrementally learns the tokens in the URL which are useful for classification.

Even though genre identification from the URL of a web page is an important problem and plays a key role in solving Information Retrieval problems like focused crawling and WPC, there is 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 [14] [13]. One cannot always afford to have these resources. Many works like [14] report their results on toy datasets and hence are not scalable to be used

in search engines like *Sandhan*[4]. In this work we build a URL based genre identification system on a huge dataset using minimal resources and training.

3 Experimental Dataset

To start with, we have a set of approximately 3000 seed URLs collected and tagged manually into tourism, health and miscellaneous genres as shown in table 1. With these seed URLs we crawl the web till a depth of 3. After eliminating other language pages and pages which are not properly parsed (about 25000 in number), we are left with 94995 **Hindi** pages.

Since such a huge number of pages cannot be tagged manually, we use the WPC algorithm based on the parsed text of the web pages as used in [3]. The web pages corresponding to the URLs mentioned in table 1 are used to train this algorithm. The 10 fold cross validation accuracy of this algorithm is 81.89% . Using this we tag the 94995 Hindi web pages. This is **assumed** to be our golden data and is used to evaluate our approach. Note that in our golden data we have preserved the order in which the pages were crawled. An interesting observation is that more than 25% of web pages are not used because of parsing issues.

Table 1. Statistics of data collected

Genre	No of Web Pages
Tourism	885
Health	978
Miscellaneous	1354

4 System Architecture

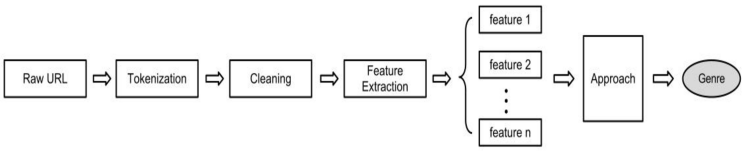

Fig. 1. System Architecture

[4] http://www.clia.iitb.ac.in:8080/sandhan-prsg/locale.jsp?en

Figure 1 describes the overall system architecture. A given URL is initially tokenized using common separators like full stop, hyphen, etc. The tokens which do not contain genre specific information are removed in the cleaning phase. In the feature extraction module various features like words and n-grams (4 to 8 grams) are extracted as used by Anastácio et al. [10]. Finally, the approach consumes these features to output the genre of the URL. The following example explains the system architecture in brief.

Raw URL: http://origin-hindi.webdunia.com/tourism.html
Tokenization: http origin hindi webdunia com tourism html
Cleaning: origin hindi webdunia tourism
Feature Extraction: Words - {origin, hindi, webdunia, tourism}
4-Grams - {orig, rigi, igin, ginh, inhi, nhin, hind, indi,.....}
5-Grams - {origi, rigin, ginhi, inhin, nhind, hindi,.....}
6-Grams - {origin, riginh, iginhi, ginhin, inhind, nhindi,....}
7-Grams - {originh, riginhi, iginhin, ginhind, inhindi,.....}
8-Grams - {originhi, riginhin, iginhind, ginhindi,.....}
All-Grams - Set of all grams (4 to 8) mentioned above

In the following section we propose three approaches to solve the problem of identifying the genre of a web page using only its URL, namely list based, naive Bayes and incremental naive Bayes.

5 Proposed Approaches

5.1 List Based Approach

This is a well known method which uses a list of tokens/words that are considered to be indicative of the genre. The list is matched against the tokens of the URL under consideration. For the URL to belong to that genre, at least one of its tokens should be present in the list. We now describe methods to prepare such a list for Hindi.

Manual Collection: A list of words (List 1) belonging to a specific genre are manually collected. For example, the list for tourism genre will contain words like travel, tour, flight, etc. and the list for health genre will contain words like health, disease, care, vitamin etc. in health genre. A glaringly obvious disadvantage of this approach is that it is effort intensive and requires genre/domain expertise.

Via an External Corpus: A corpus of sentences belonging to tourism and health genres are publicly available [5] for English and various Indian languages.

[5] `www.tdil-dc.in/index.php?option=com_up-download&view=download&lang=en&limitstart=70`

Figure 2 part (b-1) depicts the creation of the list using the genre specific corpus of English sentences. Nouns are extracted from the English sentences to give rise to List 2. This list is further augmented using WordNet which gives rise to List 3.
It is important to note that URL tokens for Hindi web pages might contain Hindi words transliterated to English. This can not be captured by the earlier lists. So we use a corpus of Hindi sentences. Figure 2 part (b-2) shows the creation of this list. Nouns are extracted from the Hindi sentences and transliterated to English using a proprietary transliteration engine. This gives rise to List 4. This list is further augmented using IndoWordNet which gives rise to List 5.

Via a Retrieval System: The previous approach depends on the presence of a tagged sentence corpus specific to a particular genre. This is a severe drawback since such a corpus may not be available for many genres. To overcome this limitation, we explore the use of the web to get a list of words specific to almost any genre. We propose the use of Wikipedia categories and search engines for this purpose. Wikipedia is an online collaborative knowledge sharing system. All pages in Wikipedia are assigned to at least one category. Each of these may be subdivided into several subcategories. Since Wikipedia content is constantly updated by a diverse audience, almost every genre will have its representation in Wikipedia as a category. We use subcategory information because of the wide coverage and diversity it offers within the genre.
Figure 2 part (c) shows how a genre specific list is created via a retrieval system. Given a genre, the corresponding Wikipedia category and all of its subcategories are extracted. Keywords are manually crafted for each of these subcategories and queried on to a search engine. Features are extracted from the top 'k' URLs to get the genre specific list of tokens.
There are several disadvantages of using the list based approach. Firstly, it is blind to the number of matching tokens in the URL. Moreover, one might not be able to capture all the genre specific tokens exhaustively. Further, attempts at increasing the list size to include more genre specific tokens may introduce noise, that is, generic words or words specific to other genres might creep in.

5.2 Naive Bayes Approach

In this method we train a naive Bayes algorithm to learn the patterns in URLs of a specific genre. Naive Bayes is the most popular text classification algorithm and is known to be accurate despite the independence assumption that it makes. Also, the training time of naive Bayes is much less when compared to other machine learning algorithms.

5.3 Incremental Naive Bayes

The naive Bayes approach is limited by the amount of training data. Moreover, the prior probabilities used in the algorithm depend on the

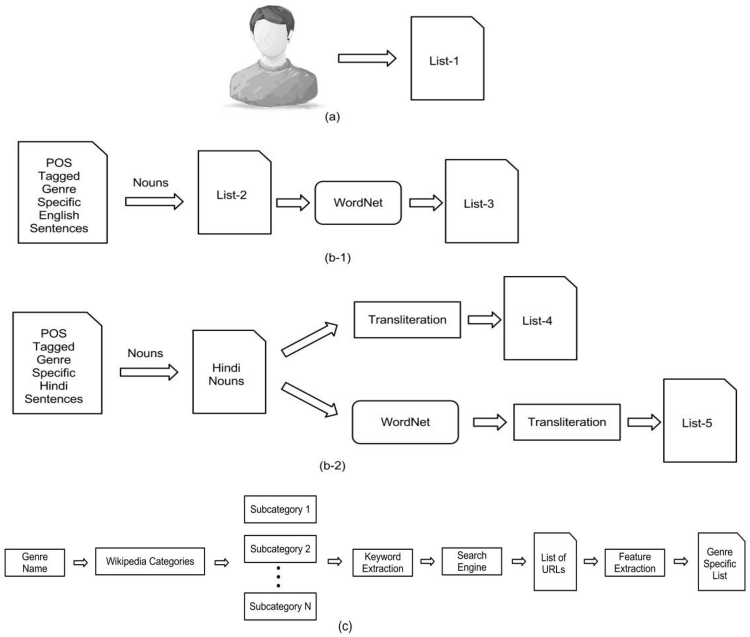

Fig. 2. List Based Approaches (a) Manual Collection (b-1) Via an External English corpus (b-2) Via an External Hindi corpus (c) Via a Retrieval System

distribution of previously observed training data. This introduces a classification bias towards the dominant class. To counter both these, we use an incremental version of the naive Bayes algorithm along with uniform prior probabilities. This has several advantages. This method incorporates the knowledge from all previously observed URLs and constantly improves as more and more URLs are observed. The uniform prior probabilities ensure that the current classification task is not affected by previous class distribution of training data.
Specifically in the context of focused crawling, where URL based genre identification is of significant use, the incremental naive Bayes algorithm is an ideal approach. In such a scenario, this algorithm can continue to train on newly crawled URLs and improve its accuracy. All the previous approaches would be unable to take advantage of the information from these newly crawled URLs.

6 Experimental Setup

In this work, we experiment with Hindi language tourism and health URLs. For the manual list based approach, a total of 29 words were selected from each genre. In case of the list based approach via a retrieval

system, tokens of less statistical significance were not considered to be representative of the genre, and hence, eliminated. Similarly, tokens like wiki, pedia, etc. which were found in both genres were also removed. In the naive Bayes approach, we use 60% of the annotated data for training and report the results for the remaining 40% testing data.

7 Evaluation Metrics

We evaluate our results using different metrics like classification accuracy and rate of change of classification accuracy wherever applicable.

$$Classification\ Accuracy = \frac{\#\ of\ correctly\ classified\ samples}{Total\ \#\ of\ samples} \quad (1)$$

8 Results and Analysis

Table 2. Accuracy of List Based Approach

Genre	Tourism	Health
List 1	0.766	0.753
List 2	0.359	0.284
List 3	0.339	0.276
List 4	0.589	0.331
List 5	0.362	0.270

Table 3. Accuracy of List Based Via a Retrieval System

Genre	Tourism	Health
4 grams	0.399	0.344
5 grams	0.483	0.477
6 grams	0.575	0.562
7 grams	0.632	0.607
8 grams	0.657	0.633

Table 2 shows the results for the list based approach using manual collection and external corpus respectively. Considering that a certain amount of manual intelligence was used, these results are not satisfactory, especially for the health genre. There may be several reasons for this. While creating these lists, n-grams were not used. The manual way of collecting cannot be exhaustive, while the corpus based approach cannot be accurate due to the presence of noise. Both of them are incapable of capturing URL vocabulary. On the other hand, the retrieval based system

Table 4. Accuracy of Naive Bayes

Genre	Tourism	Health
Words	0.848	0.446
4 grams	0.845	0.363
5 grams	0.845	0.463
6 grams	0.844	0.433
7 grams	0.839	0.456
8 grams	0.846	0.509

Table 5. Accuracy of Incremental Naive Bayes Approach

Genre	Tourism	Health
Words	0.849	0.862
4 grams	0.849	0.851
5 grams	0.858	0.865
6 grams	0.860	0.873
7 grams	0.860	0.875
8 grams	**0.858**	**0.873**
All grams	0.856	0.867

uses n-gram features and captures URL vocabulary. As shown in table 3, it outperforms both manual and corpus based approaches.

From table 4, we can see that even after giving 60% of the labeled set as training data, the naive Bayes approach does not perform satisfactorily on the health genre. Even though this system performs well on the tourism genre, it is not scalable because it uses an enormous amount of data for training. Moreover, if the amount of training data is less, the performance is sure to reduce. For the naive Bayes to perform well, the training and testing data must come from the same distribution. It is highly unlikely that the fixed amount of training data that it gets is a good representation of the entire genre.

The results for the incremental naive Bayes approach are shown in table 5. While doing experiments we use various feature spaces which include words, n-grams and all grams. We have seen that for this particular problem, n-gram features perform better than words. This is due to the fact that URLs are generally noisy, and using n-grams handles noise, spell variations and short forms in a much more efficient way.

The rate of change of accuracy curves for table 5 are shown in figures 3 and 4 for tourism and health genres respectively.

Note that Y-axis shows the classification accuracy which varies between 0 and 1. But we have showed only the variations between 0.83 to 0.89. This is because the learning curve is very fast for the first 100 URLs, and if we try to show the classification accuracy between 0 and 1, the graph becomes skewed. We can see that 5 out of 6 n-gram features work better than word features.

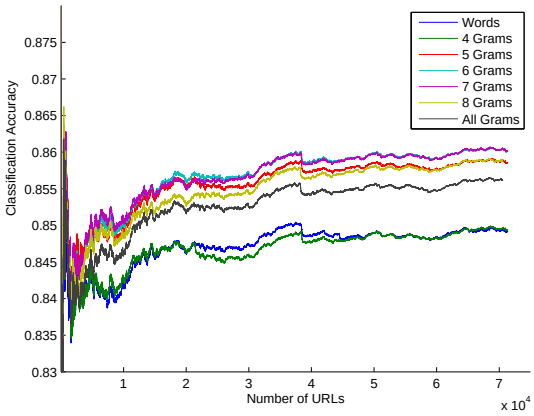

Fig. 3. Performance comparison of different features in Tourism genre

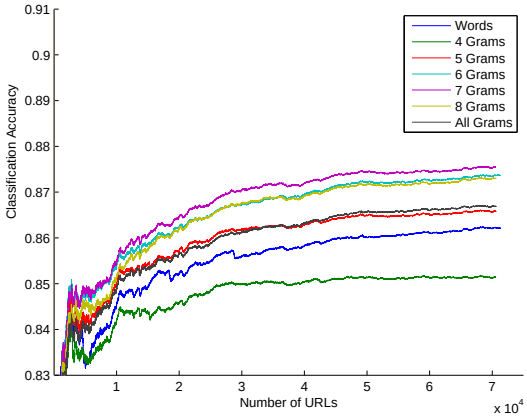

Fig. 4. Performance comparison of different features in Health genre

9 Conclusion and Future Work

In this work we have built a web page classification algorithm for Hindi tourism and health genres using only URLs of the web pages. All the experiments were done on a huge dataset containing thousands of pages. We have specifically shown its usage in the context of a focused crawler. To the best of our knowledge, there is no prior work reported for this problem in Indian languages. We are the first ones to provide a working solution for this. We propose three different kinds of approaches, from the very basic to advanced, to solve this problem: list based, naive Bayes and incremental naive Bayes approaches. The incremental naive Bayes gave the best results. Our proposed approaches are generic and can be easily extended to other languages and genres.
In future we want to try a hybrid approach which uses all the three approaches proposed in this work. We also want to conduct experiments for various other Indian languages and extend it to other genres like sports, entertainment etc.

References

1. Sharma, A., T.D.: Google sees india web explosion (2012)
2. Bhattacharyya, P.: Indowordnet. (2010)
3. Priyatam, P., Vaddepally, S., Varma, V.: Domain specific search in indian languages. In: Proceedings of the first workshop on Information and knowledge management for developing region, ACM (2012) 23–30
4. Baykan, E., Henzinger, M., Marian, L., Weber, I.: Purely url-based topic classification. In: Proceedings of the 18th international conference on World wide web, ACM (2009) 1109–1110
5. Baykan, E., Henzinger, M., Weber, I.: Web page language identification based on urls. Proceedings of the VLDB Endowment **1**(1) (2008) 176–187
6. Kan, M.Y., T.H.: Fast webpage classification using url features. In: Proceedings of the 14th ACM international conference on Information and knowledge management, ACM (2005) 325–326
7. Shih, L.K., K.D.: Using urls and table layout for web classification tasks. In: Proceedings of the 13th international conference on World Wide Web, ACM (2004) 193–202
8. Kan, M.: Web page classification without the web page. In: Proceedings of the 13th international World Wide Web conference on Alternate track papers & posters, ACM (2004) 262–263
9. Hernndez, I., R.C.R.D.C.R.: A statistical approach to url-based web page clustering. In: Proceedings of the international World Wide Web conference, ACM (2012)
10. Anastácio, I., Martins, B., Calado, P.: Classifying documents according to locational relevance. Progress in Artificial Intelligence (2009) 598–609
11. Ma, J., Saul, L., Savage, S., Voelker, G.: Identifying suspicious urls: an application of large-scale online learning. In: Proceedings of the

26th Annual International Conference on Machine Learning, ACM (2009) 681–688
12. Toyoda, Y.C.M., k.M.: Topic classification of spam host based on urls. (2010)
13. Kołcz, A., H.G.S.J.: Topical host reputation for lightweight url classification. Technical report
14. Abramson, M., A.D.: Whats in a url? genre classification from urls. In: Proceedings of Association for the Advancement of Artificial Intelligence (www.aaai.org), AAAI (2012)
15. Charu, C.A, A.G.F.Y.P.: Intelligent crawling on the world wide web with arbitrary predicates. In: Conference proceedings of World Wide Web 2010. (2010)