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]

(1) Search and Information Extraction Lab, IIIT-Hyderabad, India

(2) Tata Research Development and Design Centre, TCS-Pune, India

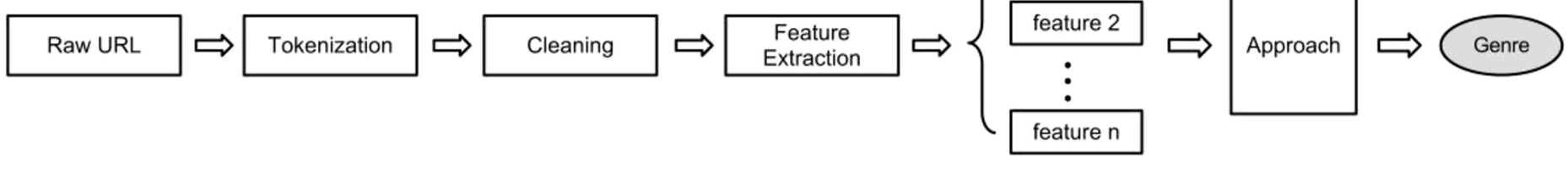
When?

- Classification speed must be high.
- Content filtering to be done before downloaded.
- Page containing images or non-standard encodings.
- Annotation on hyperlinks in a personalized web browser.
- Focused crawler wants to infer the topic of a target page.
- Language of the page needs to be identified.

Papers	Summery	Features used
Baykan et. al. [WWW 2009]	Classify web pages into 15 topics (sports, news etc.) using binary classifiers.	Use words + n-grams from the first two levels of the ODP hierarchy
Kan et. al. [CIKM 2005]	Web page classification using URL features	Position, length and sequence of tokens in a URL
Shih et. al. [WWW 2004]	Web page classification for content recommendation and ad blocking	URL tokens and the their placement in the referring page (HTML tree structure)
Kan et. al. [WWW 2004]	New segmentation techniques introduced using an information theoretic method.	Tokens obtained using information content of other partitions over and above the non-alphabetic characters.
Anastacio et al. [PAI 2009]	Categorizing documents according to their implicit locational relevance.	URL n-grams with assign weights according to the TF-IDF scheme.
Abramson et al. [AAAI 2012]	They build classifiers (Naive Bayes and SVM) using Santini and Syr7 datasets.	Syntactic and semantic style features, POS tags, punctuations and special characters

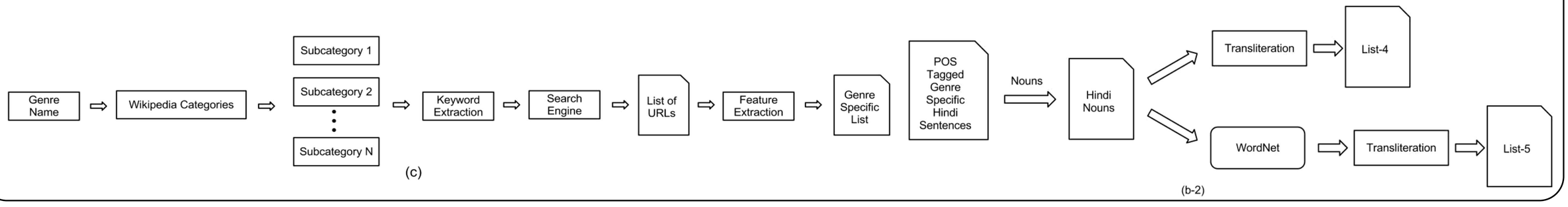
Experimental dataset

Genre	# of pages
Tourism	885
Health	978
Misc.	1354



List based approaches

(a)



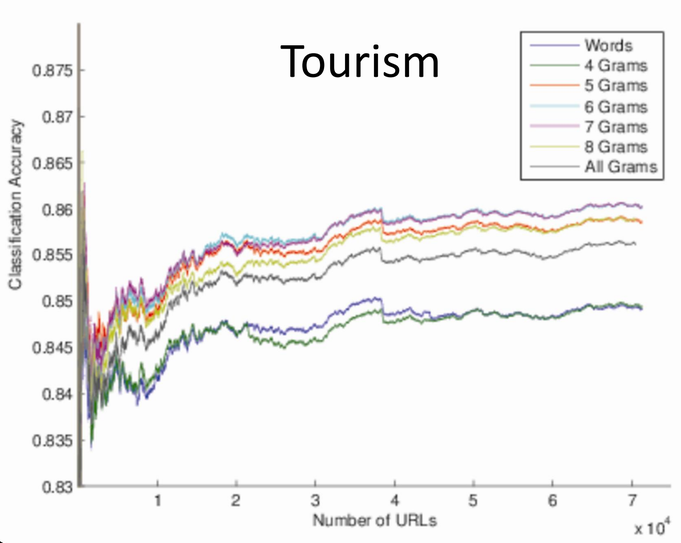
Accuracy - i) List based, ii) List based via information retrieval and iii) Naïve Bayes

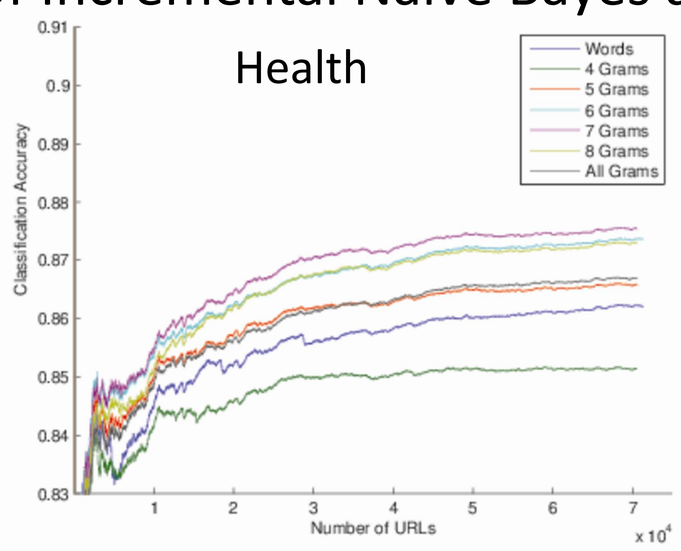
Genre	Tourism	Health
List 1	0.766	0.753
List 2	0.359	0.284
List 3	0.339	0.276
List 4	0.589	0.331
List 5	0.362	0.27

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

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

Accuracy of Incremental Naïve 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