Demographic Prediction Based On User's Browser History

Sai Kasyap Kamaraju

September 2018

1 Data parameters

Required data features for Gender and Age Prediction.

- Web pages (URLs) = $(P_1, P_2, ...P_n)$
- times = $(t_1, t_2, ...t_n)$
- Content of web pages
- Access times and order of web-pages in sequential.
- Age of the user (Definitely required for Age prediction).

User ID	Gender	Age	URL (multiple)	URL Info	Access times in order

Features that can be considered but not necessarily required....

- Sectioning on the basis of HTML Tags
- days of week, intensity, frequency of the user
- browser, OS , time active on page . etc.

2 Features

2.1 Category Based

Map web pages to standard categories , say 12.

- Alternate way is to use keywords and entities (sub cateories).
- Category scheme search ; eg : WADA in case of vietnam paper.
- Categories : ;$(C_1, C_2, ...C_n)$ Normalize to 1.
- Subcategories, say 130.
- Cons : No training data available to use a classifier (TF-IDF, Naive Bayes).

2.2 Topic Based Features

: Using LDA/ word2vec on textual Content of web pages and use topics to create additional features. [2]

- Form a corpus, use LDA models and Gibbs sampling Algorithm.
- Select topics with highest probability from the topic distribution inferred.
- Each page is mapped in to 'K' topics (K is prechosen), number of times the user has accused these 'K' topics (Normalize to 1).

2.3 Time Features

- Use one hour intervals to represent time ([0,1,2..,23]).
- For given user, count number of times the user has clicked a page in each of 24 hour times intervals(End result : Normalized vector of 24 elements).

2.4 Sequential Features

Order of viewing is also influential.

- Extract all K-grams from pages and categories; where each K-gram is of sub sequence length k(pre chosen).
- For k-gram starting at position i $(C_i, C_{i+1}, ...C_{i+k-1})$
- Don't consider k-grams if User's history is recorded from multiple sessions.(only single sessions are counted).
- For 12 Categories u will have 12^k k-grams.
- Use the mutual info between K-grams and Gender and select which have Max. Correlation.

2.5 Content based features

- Remove stop words from the URL content.
- select words based on DG /IG. [1]

2.6 Combining Features

- Topic +Time+sequential based proved equivalently effective.
- Use ML algorithms such as SVM, RF, XG-BOOST over combination of features .

3 Age Prediction :

Apriori Algorithm , a data mining and ML based Algo is used for Age prediction [3].

References

[1] Jian Hu, Hua-Jun Zeng, Cheng Niu and Zheng Chen. *Demographic Prediction Based on User's Browsing Behavior*. Microsoft Research Asia, 2007.

[2] Do Viet Phuong and Tu Minh Phuong *Gender Prediction Using Browsing History*. Vietnam.

[3] Chinmay Prakash Swami, Sumit and Prasad. *Detecting the Age of a Person through Web browsing Patterns* International Journal of Computer Applications, may 2015.