Development of a classifier to measure the commerciality of Web Documents

UNIVERSITÄT DUISBURG ESSEN

Open-Minded

Kiran Mishra, University of Duisburg-Essen

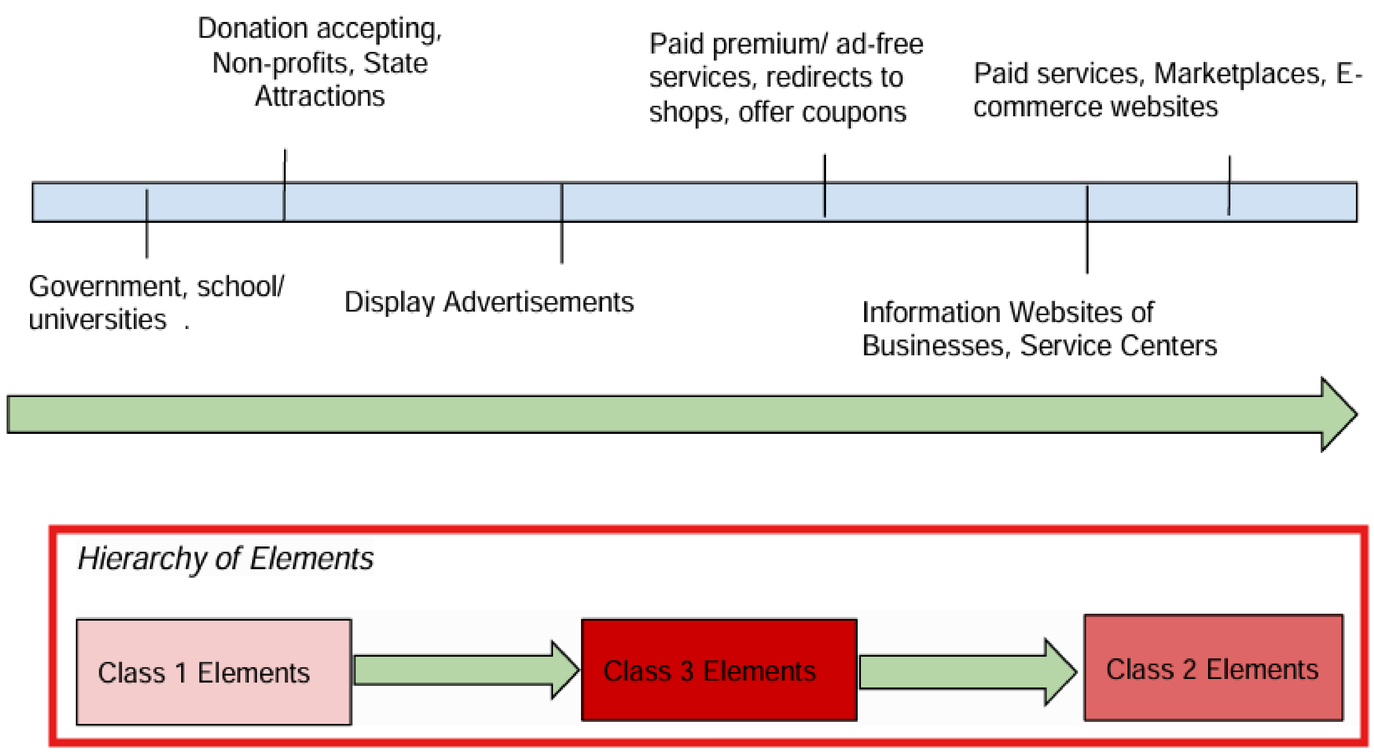
Introduction

Internet has given the power and freedom of having the access to infinite knowledge at fingertips, but getting person-dependent relevant information is a task. That is where Search Engine Ranking comes into play, but it also takes multiple features and factors into account e.g. location, search history. Along with relevance, what also matters for the search engines is the financial gain, which they achieve majorly by advertisements. (Beckett, 1996; Lewandowski, 2023; Finn et al. 2001)

This thesis analyzed how their commercial intent i.e. commerciality affects search engine ranking and tried out 3 approaches of classifying accordingly.

Method

1. Custom Dataset(image and HTML) creation:-
   - Query Words: ambiguous, so as to not bias the classifier(s) towards the query words
   - Screenshots and Source code collection
2. Classification Schema definition – 3 classes



3. Classifier Development
   - CNN
   - Object Detection - Roboflow
   - HTML Based : Contains more information about the webpage, can be found in the raw HTML markup file. (Gupta et al., 2005)

Results

	CNN	Object Detection with Roboflow	HTML- text classification
Accuracy (in %)	65.21	47.1	72.29

CNN	• Positioning and formatting dependent
Object Detection	• Non-commercial websites do not have any clear visible indicators • Works decently with defined objects.
HTML Based	• Most Efficient • Identifies the patterns of links and URLs well

- Confirms that commerciality alone does not determine ranking; both commercial and non-commercial websites appear on search engine results pages (SERP).
  - Other ranking factors include site popularity, content freshness, and geographical location. (Lewandowski, 2023)
  - Non/less commercial results were not displaced only due to commercial intent.

References
Lewandowski, D. (2023). Understanding Search Engines. Springer Nature.,
Beckett, J. (1996). The Internet phenomenon. Engineering Science & Education Journal, 5(3), 102-104.
Finn, A., Kushmerick, N., & Smyth, B. (2002). Genre classification and domain transfer for information filtering. In Advances in Information Retrieval: 24th BCS-IRSG European Colloquium on IR Research Glasgow, UK, March 25–27, 2002 Proceedings 24 (pp. 353 362). Springer Berlin Heidelberg.
Gupta, S., Kaiser, G. E., Grimm, P., Chiang, M. F., & Starren, J. (2005). Automating content extraction of html documents. World Wide Web, 8, 179-224

Contact: kiran.mishra@rwth-aachen.de