On latent domain-specific textual preferences in solving Internet-based generic tasks among graduates/young professionals from three domains

Alexander Mehler¹ Maxim Konca¹ Marie-Theres Nagel² Andy Lücking¹ Olga Zlatkin-Troitschanskaia²

¹Goethe-Universität Frankfurt, Text Technology Lab
²Johannes Gutenberg-Universität Mainz

GEBF, 09–11. 03. 2022
Critical Online Reasoning

Critical thinking wrt. online information:
(i) online information acquisition
(ii) critical information evaluation
(iii) reasoning based on evidence, argumentation, and synthesis

Online study: e-bike and health (German, browser logged)
Participants: Graduates/young professionals in medicine, law, and teaching (domains)

---

Questions

1. domain difference in search behavior?
2. domain difference among consulted websites?

- @1: search terms and types of consulted websites
- @2: text-technological classification of websites (text features)
green = teaching, red = law, purple = med.

- mutual search term overlap

→ Observe the large proportion of domain-specific search terms
# @1: Search term examples (selection)

<table>
<thead>
<tr>
<th>Shared</th>
<th>Only Medicine</th>
<th>Only Teacher</th>
<th>Only Law</th>
</tr>
</thead>
<tbody>
<tr>
<td>förderung (aid)</td>
<td>adhärenz (adherence)</td>
<td>gewerblich (commercial)</td>
<td>ärzteblatt ![1]</td>
</tr>
<tr>
<td>studie (study</td>
<td>einfluss (influence)</td>
<td>contra</td>
<td>staatsexamen (state examin.)</td>
</tr>
<tr>
<td>gesundheit (health)</td>
<td>wikipedia</td>
<td>kaufen (buy)</td>
<td>obesity (“fatness”)</td>
</tr>
<tr>
<td>e-bike (various spellings)</td>
<td>pulitzer</td>
<td>mhh (med. univ.)</td>
<td>gesundheitsministerium (health ministry)</td>
</tr>
<tr>
<td>gesundheitsfördernd (health enhancing)</td>
<td>quarks.de</td>
<td>nutrition</td>
<td>bias</td>
</tr>
<tr>
<td>google</td>
<td>kosten (cost)</td>
<td>activity</td>
<td>daily</td>
</tr>
<tr>
<td>health</td>
<td>jmir</td>
<td>testberichte (test reports)</td>
<td>empfehlung (recommend.)</td>
</tr>
<tr>
<td></td>
<td>...</td>
<td>ranking</td>
<td>...</td>
</tr>
</tbody>
</table>
shared search terms pretty directly reflect the task
other ones draw on common ground develop in the domains, but also show individual background knowledge
different search terms lead to different results and hence visited websites, but which ones?
Mainz (prev. talk): type and reliability
- 2 raters
- kappa = 0.76, Krippendorff’s alpha = 0.759
  $0.6 < \alpha \leq 0.8$ substantial agreement
  $\alpha > 0.8$ near-perfect agreement

Wiki-based
1. Wikidata classification (usually manyfold)
2. Wikipedia (“X is a(n) …”)
3. free web search (“X is a(n) …”)

⇒ our focus today!

examples: search engine, online shop, university medical database, television program, business magazin, open-acess journal, blog, portal, daily newspaper, scientific journal, online dictionary, trade magazin, private supplementary health insurance, manufacturer of fast e-bikes, …
@2: Search spaces and paths

color: type of website (Google is yellow), size: duration of visit, edge: link path

teacher trainee  medicine  law
Of course, a search engine is the center of an online search. But two retrieval patterns become visible across domains:

- follow hyperlinks
- strictly toggle between search engine and retrieved websites
INDIVIDUAL EXAMPLES TEACHER TRAINEE
INDIVIDUAL EXAMPLES MEDICINE
INDIVIDUAL EXAMPLES LAW
Main question: do the websites differ in terms of text characteristics?

A profile for each text/website is generated from a set of features.

- General features
- Lexical features
- Syntactic features
- Lexical features (HTML)
- Syntactic features (HTML)
- Lexical cohesion features

> 300 features in total

Concrete features: count features (we see a more abstract one below)
@2: TEXT SEPARABILITY AND SEPARABILITY FORCE OF FEATURES

text similarity

feature similarity (green square: HTML)
The websites which are visited only by a single domain can indeed be separated by a number of features. The top ones include:

**teacher vs. med** (93 feat $p \leq 0.05$)
- f$_{34}$ POS 0.000000002
- lmu syn. 0.000000005
- lG syn. 0.000000018
- bsesim cohs. 0.000000063
- lnHTMLbr HTML 0.000000169
- bcmu syn. 0.000000218

**law vs. teacher** (57 feat $p \leq 0.05$)
- G STO 0.0000391
- ttr STO 0.00166
- f$_{9}$ STO 0.00182
- h STO 0.00361
- btgest cohs. 0.00383

**law vs. med** (108 feat $p \leq 0.05$)
- lnHTMLbr HTML 0.000000092
- bsesim cohs 0.000000096
- Lmu syn. 0.000000155
- f$_{34}$ POS 0.000000161
- cG syn. 0.000000321
among the best separating kinds of features are
- Statistical text organization (STO; e.g. freqs., type–token ratio, entropy, …)
- cohesion (BERT-based, see below)
- syntax (dependency-related)
- HTML

@2: Distance from intersection

Do the domains differ wrt. to the similarity of the websites visited exclusively to those visited by all?

(Note that negative distance correlations are due to Gaussian smoothing)
Q: critical online reasoning: domain-specific or general?
- domain differences in search behavior
- domain differences in quantitative linguistics profiles of consulted websites
- new computational linguistics method for assessing text characteristics in educational contexts
- follow-ups: correlate text characteristics and
  - task success / polarity
  - website credibility
  - domain prediction / website recommendation
  - ...