Facilitating Access to Health Web Pages with Different Language Complexity Levels

Marco Alfano[1,4] [a], Biagio Lenzitti[2] [b], Davide Taibi[3] [c] and Markus Helfert[1] [d]

[1]*Lero, School of Computing, Dublin City University, Glasnevin Campus, Dublin, Ireland*
[2]*Dipartimento di Matematica e Informatica, Università di Palermo, Palermo, Italy*
[3]*Istituto per le Tecnologie Didattiche, Consiglio Nazionale delle Ricerche, Palermo, Italy*
[4]*Anghelos Centro Studi sulla Comunicazione, Palermo, Italy*
marco.alfano@dcu.ie, biagio.lenzitti@unipa.it, davide.taibi@itd.cnr.it, markus.helfert@dcu.ie

Keywords: e-Health, Health Information Seeking, User Requirements, Language Complexity, Structured Data on the Web.

Abstract: The number of people looking for health information on the Internet is constantly growing. When searching for health information, different types of users, such as patients, clinicians or medical researchers, have different needs and should easily find the information they are looking for based on their specific requirements. However, generic search engines do not make any distinction among the users and, often, overload them with the provided amount of information. On the other hand, specific search engines mostly work on medical literature and specialized web sites are often not free and contain focused information built by hand. This paper presents a method to facilitate the search of health information on the web so that users can easily and quickly find information based on their specific requirements. In particular, it allows different types of users to find health web pages with required language complexity levels. To this end, we first use the structured data contained in the web to classify health web pages based on different audience types such as, patients, clinicians and medical researchers. Next, we evaluate the language complexity levels of the different web pages. Finally, we propose a mapping between the language complexity levels and the different audience types that allows us to provide different types of users, e.g., experts and non-experts with tailored web pages in terms of language complexity.

1 INTRODUCTION

The number of people looking for health information on the Internet has been steadily growing over the years (Taylor, 2010; Pew Research Center, 2013) even though Akerkar and Bichile (2004) argue whether looking for health information on the Internet leads to patient empowerment or to patient deception. Although looking for medical information on the Internet may present some drawbacks, such as the amount and quality of information, they show that Internet has a major influence on patients/citizens since more than 70% of them declared that the medical information found on the Internet influenced their treatment decisions. Moreover, patients/citizens found relatively easy to obtain medical information using the Internet and one third found it easier than asking their health-care professionals.

When searching for health information on the Internet, different types of users should easily find the information they are looking for based on their specific requirements (Alfano et al., 2015a; Banna et al., 2016; Pletneva et al., 2011; EDC, 2011). In fact, patients, physicians and medical researchers have diverse needs and bring different levels of reading ability and prior knowledge together with a different vocabulary (Seedorff and Peterson, 2013; Zielstorff, 2003). However, generic search engines (like Google©, Bing© or Yahoo©) work on the whole web but make generic searches often overloading the user

---

[a] https://orcid.org/0000-0001-7200-9547
[b] https://orcid.org/0000-0003-2664-7788
[c] https://orcid.org/0000-0002-0785-6771
[d] https://orcid.org/0000-0001-6546-6408

with the provided amount of information. Moreover, they are not able to provide specific information to different types of users. On the other hand, specific search engines, such as PubMed[1] or Quertle[2], mostly work on medical literature. They provide extracts from medical journals that are mainly useful for medical researchers and experts but not for non-experts. Moreover, they do not consider all the information contained in the web that may provide additional insights. Another source of information comes from the specialized web sites oriented either to non-experts (e.g., WebMD [3], Healthline [4] or MedlinePlus [5]) or to medical professionals (e.g., Health on Net Foundation Select [6], Translating research into practice[7] or MDConsult[8]). Those sites contain very focused information but are mainly built by hand and then miss all the huge amount of information that is available on the web. Moreover, they are often not free.

A question then arises: Is it possible to facilitate the search of health/medical information on the web so that users can easily and quickly find information based on their specific requirements? In this paper, we provide a first answer to this question by presenting a system that allows different types of users to find web pages with proper language complexity levels. To this end, we first present a short survey that shows as a growing number of different users use search engines to look for health information on the web. We then use the structured data present in the web to classify health web pages based on different audience types such as patients, clinicians and medical researchers. Next, we present the results of some experiments to evaluate the language complexity levels of the different web pages and propose a mapping between the language complexity levels and the different audience types that allows us to provide users with proper web pages in terms of language complexity.

2 SEEKING HEALTH INFORMATION ON THE INTERNET

We now present a short survey of the main characteristics related to health information seeking on the Internet, based on the following dimensions:

- Who (e.g., number of people searching for health information in the Internet)
- Where (e.g., search engines, social networks)
- When (e.g., time frequency)
- What (e.g., symptoms, pathologies, remedies, drugs)

The 'Cyberchondriacs' Harris Poll (Taylor, 2010) shows that the percentage of all US adults who search for health or medical information online has increased from 27% to 76% from 1998 to 2010. On the other hand, the 'Health Online 2013' Pew report (Pew Research Center, 2013) says that 72% of adult users in the U.S. were looking for health information online in the previous year. When asked to think about the last time they went online for health or medical information, 39% of online health seekers say they looked for information related to their own situation. Another 39% say they looked for information related to someone else's health or medical situation. An additional 15% of these internet users say they were looking both on their own and someone else's behalf.

For what concerns Europe, (Kummervold, 2008) shows a growth from 14% to 39% in the 2005-2007 period. Moreover, in 2010, national bodies reported that 52,5% of adults in Spain were looking for health content on the Internet (Instituto Nacional de Estadística, 2010) and 39% in the UK (UK national statistics, 2010).

According to (Pew Research Center, 2013), 77% of online health seekers say they began their last session at a search engine such as Google, Bing, or Yahoo. Another 13% say they began at a specialized site in health information, like WebMD. Just 2% say they started their research at a more general site like Wikipedia and an additional 1% say they started at a social network site like Facebook. According to the survey reported in (Pletneva et al., 2011), a general search engine is the most frequently used tool to look for online health information (82% of the participants said that they use a search engine always or often). Other popular sources include websites providing health information (38%) and Wikipedia or medical search tools such as HONselect and Medline Plus (37%). Forums and blogs are always or often used by 23% of the respondents and 5% use Facebook or other social networks.

The same paper affirms that Internet is the second source of information after physicians whereas

---

[1]http://www.ncbi.nlm.nih.gov/pubmed/
[2]https://quertle.com/
[3]http://www.webmd.com/
[4]http://www.healthline.com/

[5]http://www.nlm.nih.gov/medlineplus/
[6]http://www.hon.ch/
[7]http://www.tripdatabase.com/
[8]http://www.mdconsult.com/

(Keselman and Slaughter, 2007) states that Internet is the most commonly consulted resource for health information followed by conversation with health care providers and use of a medical dictionary.

Taylor (2010) shows that the percentage of US adults who often or sometimes search for health or medical information online has increased from 42% to 73% from 1998 to 2010. Moreover, 81% of health information seekers have looked for health information online in the last month and 17% have gone online to look for health information ten or more times in the last month. On average, health information seekers do this about 6 times a month. According to the survey presented in (Pletneva et al., 2011), 24% of the respondents say they were looking for health information on the Internet at least once a day (some mentioned from four to six times a day in comments) and 25% did it few times a week. Moreover, 8% did it once a week, 16% did it few times a month and 16% did it once a month.

(Pew Research Center, 2013) shows that the most searched health topics are: Specific disease or medical problem (55%), Certain medical treatment or procedure (43%), How to lose weight or how to control your weight (27%), and Health insurance, including private insurance, Medicare or Medicaid (25%). According to the survey reported in (Pletneva et al., 2011), the search activity of users is mostly focused on general health information (68%), long-term chronic diseases (59%), healthy lifestyle and nutrition (50%), short-term (up to 2 weeks) acute disease (39%), kids health (22%) and elderly health and care (19%).

The above data clearly show that there is a high number of people seeking for health information on the Internet that has been constantly increasing over the years (who). Search engines are the most used means to access medical information (where) and they are used more and more often (when) to seek information on a broad range of medical subjects (what). As a consequence, a question arises: Is it possible to facilitate the search of medical information on the web so that users can easily and quickly find the information based on their specific requirements?

To answer this question, in the next sections, we analyse what are the main user requirements when seeking for health information on the web and propose a methodology for providing users with web pages that present different language complexity levels, i.e., one of the most important user requirements.

3 USE OF STRUCTURED DATA TO CLASSIFY HEALTH WEB PAGES BASED ON AUDIENCE TYPES

When seeking for health information on the Internet different types of users have different requirements (Alfano et al., 2014; Ardito, 2013; Banna et al., 2016; Eysenbach and Köhler, 2002; Higgins et. Al, 2011; Jacobs et al., 2017). Among others, non-experts usually require the used language to be easy to understand whereas medical experts require a more technical and precise language. Both categories require the information to be trustworthy.

In order to satisfy these requirements, and mainly the one related to the language complexity, we now investigate the possibilities offered by structured data to find web pages suitable to different classes of users. To this end, we exploit the semantic information available in the World Wide Web and, in particular, the one provided by schema.org[13]. This is a very important initiative, founded by some major web players, that aims to create, maintain, and promote schemas for structured data on the Internet. In particular, schema.org defines the types and properties associated with the information included in the web pages so to expose them to search engines and make easier for people to find the 'right' web information. It is, presently, used in over ten million web sites[9], and its adoption has been investigated in previous research (Dietze, 2017).

For the scope of the present work, we consider the core schema and the *health-lifesci* extension[10] that contains 100 types, 177 properties and 147 enumeration values related to the health/medical field. In particular, we consider the *MedicalAudience*[11] type that describes the target audiences for medical web pages and *Patient*[12], *Clinician*[13] and *MedicalResearcher*[14] specific types. As reported in schema.org, a patient is any person recipient of health care services. Clinicians are medical clinicians, including practicing physicians and other medical professionals involved in clinical

---

practice, and medical researchers are professionals who make research on the medical field.

Considering the above types, we have performed an analysis based upon the schema.org information made available by the Web Data Commons initiative [15]. The Web Data Commons (WDC) (Meusel, 2014) contains all Microformat, Microdata and RDFa data extracted from the open repository of web crawl data named Common Crawl (CC)[16]. The data, released in November 2017, have been used in this work. The whole dataset contains about 3.2 billion pages, with about 38.9% of them presenting structured data.

The dataset dump available on the Web Data Commons web site that we used in our study consists of 38.7 billion RDF quads[17]. These are sequences of RDF terms in the form {s, p, o, u}, where s, p and o represent a triple consisting of subject, predicate, object and u represents the URI of the document from which the triple has been extracted. The dataset dump has been made available as compressed files and each file is around 100 MB large. Overall 8,433 files with a total size of 858 GB have been provided. From these 8,433 files we have extracted the quadruples that contain the Patient, Clinician and MedicalResearcher specific types. We have obtained, then, a subset of the WDC dataset dump that contains, for each type, the number of RDF quads reported in Table 1.

Table 1: Number of RDF Quads extracted for each specific type.

Schema.org types	RDF Quads
Patient	36,186
Clinician	15,913
MedicalResearcher	3,458

Fig. 1 presents an example of RDF quads, for the Patient subtype, extracted from Web Data Commons. It clearly shows the subject, predicate, object and URI of the quadruples. In compliance with the Open Science model, we have made available the RDF quads subsets, for the Patient, Clinician and MedicalResearcher specific types, at the address http://h-easy.lero.ie/opendata/, in order to allow other researchers to use and lead further research on these data. Fig. 2 shows an extract of five RDF quads from each subset.

Thus, by using, in turn, one of the three subsets we are able to extract web pages targeted to Patient, Clinician and MedicalResearcher types. Notice that, at this stage, we have found web pages that have been
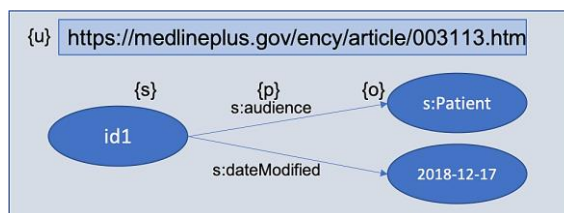

Figure 1: Example of RDF quads for the Patient subtype.

targeted to the different user types by their authors but we do not exactly know why. It could be related to the language complexity level (e.g., more or less technical) or to the treated subject (e.g., pathology symptoms and remedies, for patients, or deepening aspects, for medical researchers), or to something else. In the next section, we will map the language complexity levels to the different audience types so to be able to provide users with web pages related to their specific requirements.

4 MAPPING LANGUAGE COMPLEXITY USER REQUIREMENT TO AUDIENCE TYPES

As seen above, users (mainly non-experts) have different requirements when searching for health information on the web. In particular, one of the most important requirement for non-expert health information seekers is that the language used in the web pages must be easy to understand. On the opposite, medical experts require that the info they are looking for presents a proper technical and rigorous terminology.

We then consider two classes of users:

- Non experts (e.g., patients or citizens);
- Experts (e.g., physicians or medical researchers).

We have used the three subsets presented in the previous section, Patient, Clinician, and MedicalResearcher, and, for each quadruple, we have analysed the related web page in order to estimate its language complexity. To this end, we have evaluated the 'term familiarity index', as described in (Kloehn, N. et al., 2018; Leroy et al., 2012) of the English and non-empty web pages (around 50% of the total). In particular, for each web page, we have computed the term familiarity of each word by using the number of results provided by the Google search engine and then

Subject	Predicate	Object	URI
<https://medlineplus.gov/ency/article/001199.htm>	<http://schema.org/MedicalWebPage/audience>	<http://schema.org/Patient>	<https://medlineplus.gov/ency/article/001199.htm>
<https://medlineplus.gov/ency/article/003983.htm>	<http://schema.org/MedicalWebPage/audience>	<http://schema.org/Patient>	<https://medlineplus.gov/ency/article/003983.htm>
_:node49375316dc61759032984f402ff140	<http://schema.org/MedicalWebPage/audience>	<http://schema.org/Patient>	<https://www.healthgrades.com/group-directory/tn-tennessee/memphis>
_:node29c2dc857960d3506979ce6e9e9fcab	<http://health-lifesci.schema.org/MedicalWebPage/audience>	<http://schema.org/Patient>	<https://healthjoy.com/health-plan/coverage/chicago-il-60654/dentist/PPO-all/>
_:nodea935f7bf8568d9be197caabd279ed880	<http://schema.org/MedicalWebPage/audience>	<https://schema.org/Patient>	<https://www.ready2smile.com/temecula-ca/alternative-to-braces-for-teens/>

(a)

Subject	Predicate	Object	URI
_:nodead17e982643d4ff761893dc8d21a8a	<https://schema.org/MedicalWebPage/audience>	<https://schema.org/Clinician>	<http://www.fpnotebook.com/OB/Fetus/FtlBrdycrd.htm>
_:nodebbeb4aa9fdaa7c6bbbf84ef58572be	<https://schema.org/MedicalWebPage/audience>	<https://schema.org/Clinician>	<http://www.fpnotebook.com/ENT/Lip/AnglrStmts.htm>
_:node5ac15d6f84ffa289c1b2b15e24ce99c	<https://schema.org/MedicalWebPage/audience>	<https://schema.org/Clinician>	<http://www.fpnotebook.com/Neuro/Motor/Chr.htm>
_:nodebc14ebca2a1b81633703ba9f9108660	<http://schema.org/MedicalScholarlyArticle/audience>	<http://schema.org/Clinician>	<http://docality.com/doctor/profile/1730195587/dr-azizul-hoque-md>
_:nodea943b2c2e88ea6d7216a44b8ca4d1f6	<http://schema.org/MedicalScholarlyArticle/audience>	<http://schema.org/Clinician>	<http://docality.com/doctor/profile/1730195587/dr-azizul-hoque-md>
_:nodeebf33f24f05b1ec242d1f8af45f71f42	<http://schema.org/MedicalScholarlyArticle/audience>	<http://schema.org/Clinician>	<http://docality.com/doctor/profile/1730195587/dr-azizul-hoque-md>

(b)

Subject	Predicate	Object	URI
_:node7c9bf2b3856678d871182f3b135bd0	<http://schema.org/MedicalWebPage/audience>	<http://schema.org/MedicalResearcher>	<http://www.malacards.org/card/sarcoma_synovial>
_:nodee6cdd393ab8d0a188a48568b4a6f2	<http://schema.org/MedicalWebPage/audience>	<http://schema.org/MedicalResearcher>	<http://www.malacards.org/card/vascular_disease>
_:nodec9cc1378dc9f960c4ad9f9a788af4b	<http://health-lifesci.schema.org/MedicalWebPage/audience>	<http://schema.org/MedicalResearcher>	<https://healthjoy.com/health-plan/coverage/chicago-il-60654/sports-medicine/PPO-all/>
_:node352af2c53295fea33d31d7283802098	<http://health-lifesci.schema.org/MedicalWebPage/audience>	<http://schema.org/MedicalResearcher>	<https://healthjoy.com/health-plan/coverage/chicago-il-60654/pediatrics-behavior-and-neurodevelopment/PPO-all/>
_:node275549fc96b456291c2d630b61f3b63	<http://health-lifesci.schema.org/MedicalWebPage/audience>	<http://schema.org/MedicalResearcher>	<https://healthjoy.com/health-plan/coverage/chicago-il-60654/pain-medicine/PPO-all/>

(c)

Figure 2: An extract of five RDF quads extracted from Patient (a), Clinican (b) and MedicalResearcher (c) subsets.

we have computed the term familiarity index of the page by averaging all the word familiarity indexes. This information has been stored in a database to avoid work duplication.

In particular, for each web page, we have computed and stored the total number of words, the number of unique words, the number of least common words (i.e., the number of words minus the number of words belonging to the list of the three thousand most recurrent words as represented by the "Dale-Chall Easy Word List"[18]), the average of term familiarity indexes of all words and the average of term familiarity indexes of the least common words (in order to evaluate if the probable presence of the most recurrent words in each web page could cause a bias

in the average of familiarity indexes). The results of the performed tests, for the three audience types, are available at the address http://www.math.unipa.it/simplehealth/simple2/ResSchema.php and the first six results of each audience type are shown in Fig. 3.

Next, we have computed some statistics related the term familiarity indexes of the web pages for the different target audiences (by taking into account all words and the least common ones) and we have obtained the results reported in Fig. 4 and Fig. 5.

Fig. 4 shows, for each specific type, the box plot of the average of the term familiarity indexes computed for all words. A box plot is a standardized way of displaying the distribution of data based on a five number summary ("minimum", first quartile

---

[18] https://www.usingenglish.com/resources/wordcheck/list-dale-chall+list+of+simple+words.html

ID	URL	# Distinct Words	Average Freq.DW in billions	# Total Words	Average Freq.TW in billions	# NC Distinct Words	Average Freq.NCDW in billions
1	https://tatefamilydentistry.com/...	256	5.95	502	8.51	124	2.50
2	https://midtownoms.com/corrective-jaw-surgery/...	208	5.98	446	8.30	111	2.89
5	https://midtownoms.com/implant-bone-grafting/...	239	4.90	525	8.15	148	2.34
8	https://midtownoms.com/contact-us/...	76	7.80	139	7.18	38	5.60
9	https://www.restylaneusa.com/specialist...	361	4.72	1037	8.60	233	1.98
11	https://midtownoms.com/referring-doctors/...	92	7.56	168	7.68	48	5.17

(a)

ID	URL	# Distinct Words	Average Freq.DW in billions	# Total Words	Average Freq.TW in billions	# NC Distinct Words	Average Freq.NCDW in billions
3	https://www.onlinedentalmarketing.com/targeted-dental-market...	230	7.03	471	10.35	132	3.85
4	https://www.onlinedentalmarketing.com/targeted-dental-market...	346	5.57	790	9.96	212	3.22
10	https://www.onlinedentalmarketing.com/privacy-policy/...	552	5.36	1422	10.07	337	2.92
16	https://www.onlinedentalmarketing.com/blog/...	299	6.74	664	9.31	154	3.78
18	https://www.onlinedentalmarketing.com/meet-us/...	183	6.59	312	8.28	114	3.81
20	https://www.onlinedentalmarketing.com/targeted-dental-market...	334	5.69	692	9.23	186	2.89

(b)

ID	URL	# Distinct Words	Average Freq.DW in billions	# Total Words	Average Freq.TW in billions	# NC Distinct Words	Average Freq.NCDW in billions
6	http://hcvhub.deusto.es/	122	6.71	194	8.57	77	3.19
7	http://www.malacards.org/card/geniculate_herpes_zoster...	677	2.39	2047	3.13	521	1.00
15	http://www.malacards.org/card/yaws...	2815	0.87	6994	1.82	2283	0.42
37	http://www.malacards.org/card/klippel_feil_syndrome_3_autoso...	314	3.08	1095	2.83	246	1.48
39	http://www.malacards.org/card/chorioretinitis...	2374	0.85	5592	1.67	2006	0.39
54	http://www.malacards.org/card/spindle_cell_hemangioma...	555	2.51	1597	3.29	435	1.05

(c)

Figure 3: First six test results for Patient (a), Clinican (b), and MedicalResearcher (c) audience types.

(Q1), median, third quartile (Q3), and "maximum"). Overall, the median and the first-third quartile interval of Patient is much higher of those of Clinician and MedicalResearcher that partially overlap. The outliers above the maximum mainly refer to pages that contain informative/commercial data for the different types of users and then use a simple language. The outliers below the "minimum" mainly refer to pages, such as those of the www.malacards.org domain, which indicate all three classes, as target audiences, but have a low term familiarity index clearly indicating that they should be targeted only to medical experts for what concerns the language complexity.

Fig. 5 shows, for each specific type, the box plot of the average of the term familiarity indexes computed for the least common words (as seen above). The same considerations of Figure 4 apply to

Fig. 5 but, by eliminating the most common words, the figures of Patient lower much more than the others showing, once more, as the web pages targeted to Patient are the ones tending to use the simplest language.
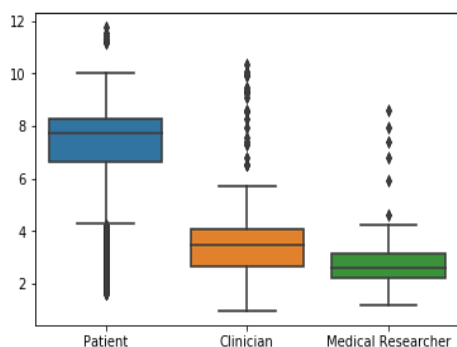

Figure 4: Box plot of the average of term familiarity indexes for all words.
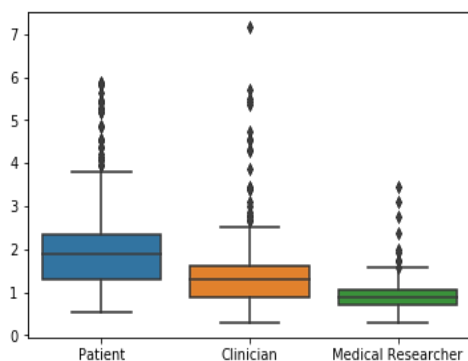

Figure 5: Box plot of the average of term familiarity indexes for the least common words.

The experimental results show that the web pages targeted to Patient, present, on average, a much higher term familiarity index and thus a simpler terminology whereas the web pages targeted to Clinician and MedicalResearcher present, on average, a lower term familiarity index and thus a more complex terminology, even though Clinician pages are a little closer to Patient pages. As a consequence, Patient pages, falling in the intervals shown in Fig. 4 and Fig. 5, can be used for the Non-expert class and Clinician/MedicalResearcher pages, falling in the intervals shown in Fig. 4 and Fig. 5, can be used for the Expert classes producing then the following mapping:

- Non-experts -> Patient
- Experts -> MedicalResearcher and Clinician

This allows us to provide different types of users with health web pages targeted to their specific language complexity requirements. Notice that the presence of structured data inside a web page can also be seen, somehow, as a guarantee of information quality even though an evaluation of the quality level of the info of a web page requires a specific analysis that is outside the scope of this work.

As a final step, we have built a navigational tool that allows to navigate among the web pages related to the different types. This navigational tool can be accessed at the address http://www.math.unipa.it/facile and Fig. 6 reports the input interface of the tool with the 'diabetes' term to be searched.

Figure 6: Input interface of the navigational tool.

Fig. 7 reports the top ten results of the navigational tool for the 'diabetes' term and for the three types: Patient, Clinician and MedicalReseracher and the related weight. The ranking of each web page is presently done by computing the page wight as follows:

- Patient

$$(Term\_Frequency * Page\_Familiarity\_Index) / Total\_Number\_Of\_Words \quad (1)$$

because we want meaningful pages (high number of occurrences of the searched item) but with the simplest language;

- Clinician and MedicalResearcher

$$(Term\_Frequency / Page\_Familiarity\_Index) / Total\_Number\_Of\_Words \quad (2)$$

because we want meaningful pages (high number of occurrences of the searched item) but with the most complex/technical language.

By examining Fig. 7 we can easily see that the top links of Patient present a high term familiarity index and belong to medlineplus.gov which is notoriously a web portal for non-experts. The top links of Clinician present a medium-low term familiarity index and belong to the fpnotebook.com web portal which acts as a medical dictionary and presents a technical language even though understandable by users with

Ranking	URL	Weight
1	http://www.northlandmed.com.au/diabetes/	146.57
2	https://medlineplus.gov/ency/article/001214.htm	130.26
3	https://medlineplus.gov/ency/article/000305.htm	111.21
4	https://medlineplus.gov/ency/article/000313.htm	108.60
5	https://medlineplus.gov/ency/patientinstructions/000328.htm	103.36
6	https://medlineplus.gov/ency/patientinstructions/000086.htm	101.76
7	https://medlineplus.gov/ency/patientinstructions/000079.htm	93.92
8	https://medlineplus.gov/ency/patientinstructions/000083.htm	89.21
9	https://medlineplus.gov/ency/patientinstructions/000085.htm	84.79
10	https://medlineplus.gov/ency/patientinstructions/000322.htm	83.67

(a)

Ranking	URL	Weight
1	http://www.fpnotebook.com/Renal/Endo/CntrlDbtsInspds.htm	109.81
2	http://www.fpnotebook.com/Endo/DM/DbtsMlts.htm	105.93
3	http://www.fpnotebook.com/Endo/DM/DbtsSckDyMngmnt.htm	98.31
4	http://www.fpnotebook.com/Endo/DM/DbtcKtcds.htm	93.70
5	http://www.fpnotebook.com/Endo/Lab/DbtsScrng.htm	90.57
6	http://www.fpnotebook.com/cv/Endo/CrnryArtryDsPrvntnInDbts.htm	89.25
7	http://www.malacards.org/card/diabetes_mellitus_transient_neonatal_3	88.97
8	http://www.malacards.org/card/maturity_onset_diabetes_of_the_young_type_10	87.78
9	http://www.fpnotebook.com/Endo/DM/DbtsMltsCntrlInHsptl.htm	86.25
10	http://www.malacards.org/card/diabetes_mellitus_6q24_related_transient_neonatal	85.04

(b)

Ranking	URL	Weight
1	http://www.malacards.org/card/gestational_diabetes_insipidus	92.40
2	http://www.malacards.org/card/diabetes_mellitus_transient_neonatal_3	88.97
3	http://www.malacards.org/card/maturity_onset_diabetes_of_the_young_type_10	87.78
4	http://www.malacards.org/card/diabetes_mellitus_6q24_related_transient_neonatal	85.04
5	http://www.malacards.org/card/monogenic_diabetes	80.11
6	http://www.malacards.org/card/diabetes_insipidus_neurohypophyseal	78.44
7	http://www.malacards.org/card/diabetes_mellitus_permanent_neonatal	78.02
8	http://www.malacards.org/card/mitochondrial_myopathy_with_diabetes	75.83
9	http://www.malacards.org/search/results/nephropathy-hypertension?retired=1	72.24
10	http://www.malacards.org/card/neonatal_diabetes_mellitus	68.71

(c)

Figure 7: Diabetes outputs for Patient (a), Clinician (b), and MedicalResearcher (c).

Table 2: Comparing Google ranking and term familiarity ranking for 'diabetes' keyword.

#	Google Ranking	Term Familiarity Ranking	Familiarity Index
1	https://medlineplus.gov/diabetes.html	https://www.nhsinform.scot/illnesses-and-conditions/diabetes/diabetes	9.43
2	https://www.webmd.com/diabetes/default.htm	https://www.nhs.uk/conditions/diabetes/	8.89
3	https://www.medicalnewstoday.com/articles/323627.php	https://www.medicalnewstoday.com/articles/323627.php	8.63
4	https://en.wikipedia.org/wiki/Diabetes_mellitus	https://www.healthline.com/health/diabetes	8.42
5	www.diabetes.org/	http://www.diabetes.org/	7.34
6	https://www.medicinenet.com/diabetes_mellitus/article.htm	https://www.mayoclinic.org/diseases-conditions/diabetes/symptoms-causes/syc-20371444	7.30
7	https://www.nhs.uk/conditions/diabetes/	https://www.medicinenet.com/diabetes_mellitus/article.htm	7.26
8	https://www.mayoclinic.org/diseases-conditions/diabetes/symptoms-causes/syc-20371444	https://medlineplus.gov/diabetes.html	7.02
9	https://www.healthline.com/health/diabetes	https://www.webmd.com/diabetes/default.htm	6.91
10	https://www.nhsinform.scot/illnesses-and-conditions/diabetes/diabetes	https://en.wikipedia.org/wiki/Diabetes_mellitus	6.01

some medical skills. The top links of MedicalResearcher present a low term familiarity index and belong to malacards.org web portal that is a human disease database and presents a very technical and complex language. Notice that, as seen above, some malacards.org pages contain all the three audience types and may appear in the other rankings because often present a high number of occurrences of the searched item. Of course, the ranking mechanism presented here is just a first proposal and needs to be refined and enriched to transform the navigational tool in a proper user-focused search engine.

The use of structured data related to the intended audience, in combination with the term familiarity of a web page, provides a method to rank web pages in terms of the complexity level of the text. Generalising this approach, the term familiarity method can be used to rank web pages even when they do not contain any specific structured data about their audience. As an example, Table 2 shows the comparison between the top ten results provided by the Google search engine for the 'diabetes' keyword (1st column) and the same set ranked according to the page term familiarity index (2nd and 3rd columns).

The applications of such ranking method can be various. To cite a few examples: a meta search engine could enrich the results of popular search engines by providing additional information about the term familiarity of each result or the results could be sorted and presented according to the requirements of expert or non-expert users.

4 CONCLUSIONS

In this work, after presenting the main characteristics related to health information seeking on the Internet, we have proposed an approach based on structured data (by using schema.org) to classify health web pages for different audience types.

Moreover, we have executed some experiments to evaluate the term familiarity indexes of different web pages and proposed a mapping between the language complexity user requirement and the different audience types. We have then presented a navigational tool that allows different users to obtain the web pages related to their language complexity requirements.

Of course, more experiments need to be executed in order to better understand the correlation between the language complexity levels and the different audience types and establish specific thresholds for what concerns the term familiarity index of a web page so that we can easily classify it as suitable to expert or non-expert. Moreover, as seen above, we need to improve the ranking mechanism of our navigational tool so that is able to provide the "correct" pages (in relation to the searched item/s) while privileging, as much as possible, the term

familiarity. Furthermore, we want to analyse how to provide users with web pages with proper language complexity levels even using pages with different complexity levels, e.g., simplifying the complex medical terminology for a non-expert (Alfano et al, 2018; Alfano et al, 2015b). Finally, we want to consider other user requirements, such as the quality of information, and analyse if and which structured data (e.g., schema.org types) could provide us, for example, with web pages which present a high quality of information.

ACKNOWLEDGEMENTS

REFERENCES

Akerkar, S., & Bichile, L., 2004. Health Information on the Internet: Patient Empowerment or Patient Deceit? *Indian Journal of Medical Sciences*, 58(8). Pp. 321-326.

Alfano, M., Lenzitti, B., and Lo Bosco, G., 2014. A web search methodology for health consumers, *Proc. of ACM International Conference on Computer Systems and Technologies (CompSysTech'14)*, Ruse, pp. 150-157.

Alfano, M., Lenzitti, B., and Lo Bosco, G., 2015a. U-MedSearch: A Meta Search Engine of Medical Content for Different Users and Learning Needs. *Proc. of International Conference on e-Learning (e-Learning'15)*, Berlin.

Alfano, M., Lenzitti, B., Lo Bosco, G., and Perticone, V., 2015b. An Automatic System for Helping Health Consumers to Understand Medical Texts, *Proc. of HEALTHINF 2015*, Lisbon, pp. 622-627.

Alfano, M., Lenzitti, B., Lo Bosco, G., and Taibi, D., 2018. Development and Practical Use of a Medical Vocabulary-Thesaurus-Dictionary for Patient Empowerment. *Proc. of ACM International Conference on Computer Systems and Technologies (CompSysTech'18)*, Ruse.

Ardito, S. C., 2013. Seeking Consumer Health Information on the Internet, 37(4), 1–5. Retrieved from http://www.infotoday.com/OnlineSearcher/Articles/Medical-Digital/Seeking-Consumer-Health-Information-on-the-Internet-90558.shtml

Banna, S., Hasan, H. & Dawson, P., 2016. Understanding the diversity of user requirements for interactive online health services. *International Journal of Healthcare Technology and Management*, 15(3).

Dietze S., Taibi D., Yu R., Barker P., d'Aquin M., 2017. Analysing and Improving Embedded Markup of Learning Resources on the Web. *Proc. of the 26th International Conference on World Wide Web Companion (WWW '17 Companion)*. International World Wide Web Conferences Steering Committee, Republic and Canton of Geneva, Switzerland, 283-292. DOI: https://doi.org/10.1145/3041021.3054160.

Eysenbach, G. & Köhler, C., 2002. How do consumers search for and appraise health information on the world wide web? Qualitative study using focus groups, usability tests, and in-depth interviews. *BMJ (Clinical research ed.)*, 324(7337), pp.573–7.

Higgins, O., Sixsmith, J., Barry, M.M., Domegan, C., 2011. A literature review on health information seeking behaviour on the web: a health consumer and health professional perspective. Stockholm: *ECDC*.

Instituto Nacional de Estadística. 2010. Encuesta sobre Equipamiento y Uso de Tecnologías de la Información y Comunicación en los hogares.

Jacobs, W., Amuta, A. O. & Jeon, K. C., 2017. Health information seeking in the digital age: An analysis of health information seeking behavior among US adults. *Cogent Social Sciences*, 3(1), pp.1–11.

Keselman, A. & Slaughter, L., 2007. Towards consumer-friendly PHRs: patients' experience with reviewing their health records. *Proc. AMIA Annual Symposium Proceedings*, pp.399–403.

Kloehn, N. et al., 2018. Improving consumer understanding of medical text: Development and validation of a new subsimplify algorithm to automatically generate term explanations in English and Spanish. *Journal of Medical Internet Research*, 20(8).

Kummervold E., Chronaki C.E., Lausen B., Prokosch H.U., 2008. eHealth Trends in Europe 2005-2007: A Population-Based Survey. *J Med Internet Res.*, Vol. 10.

Leroy, G. et al., 2012. Improving perceived and actual text difficulty for health information consumers using semi-automated methods. *AMIA Annual Symposium Proceedings*. pp.522–31.

Meusel, R., Petrovski, P., and Bizer, C. 2014. The WebDataCommons Microdata, RDFa and Microformat Dataset Series. *Proc. of the 13th International Semantic Web Conference (ISWC14)*, Springer-Verlag New York, USA, 277-292.

Pew Research Center, 2013. Health online 2013, http://www.pewinternet.org/2013/01/15/health-online-2013/.

Pletneva, N., Vargas, A. & Boyer, C., 2011. Requirements for the general public health search. *Khresmoi Public Deliverable D8.1.1*.

Seedorff, M., and Peterson, K., 2013. Incorporating Expert Terminology and Disease Risk Factors into Consumer Health Vocabularies, *Pac. Symp. Biocomp.*, pp. 421–432.

UK national statistics, 2010. Statistical bulletin: Internet Access 2010. *Office for National Statistics*. 27 Aug 2010.

Taylor, H. 2010. HI-Harris-Poll-Cyberchondrics. Harris Interactive. https://theharrispoll.com/the-latest-harris-poll-measuring-how-many-people-use-the-internet-to-look-for-information-about-health-topics-finds-that-the-numbers-continue-to-increase-the-harris-poll-first-used-the-word-cyberch/.

Zielstorff, R. D., 2003. Controlled vocabularies for consumer health", *Journ. Biomed. Inform.*, vol. 36, no. 4–5, pp. 326–333.