Author(s): Lahti, Lauri

Title: Supporting online health queries by modeling patterns of creation, modification and retrieval of medical knowledge

Year: 2016

Version: Post print

Please cite the original version:

Lahti, Lauri. 2016. Supporting online health queries by modeling patterns of creation, modification and retrieval of medical knowledge. EdMedia 2016 - World Conference on Educational Media and Technology, 28-30 June 2016, Vancouver, Canada. Proc. EdMedia 2016 - World Conference on Educational Media and Technology, 28-30 June 2016, Vancouver, Canada. 7.

Lahti, Lauri (2016d). Supporting online health queries by modeling patterns of creation, modification and retrieval of medical knowledge. Accepted to appear in Proc. EdMedia 2016 - World Conference on Educational Media and Technology, 28-30 June 2016, Vancouver, Canada. Association for the Advancement of Computing in Education (AACE), Chesapeake, VA, USA.
Open access in Aaltodoc publication archive at http://aaltodoc.aalto.fi.

Supporting online health queries by modeling patterns of creation, modification and retrieval of medical knowledge

Lauri Lahti
Department of Computer Science
Aalto University School of Science, Finland

Abstract: We evaluated properties of knowledge resources that can be used for building new semantically and behaviorally motivated resources of health guidance and clinical decision making by modeling patterns of creation, modification and retrieval of medical knowledge. We evaluated statistical properties of Wikipedia articles of general terminology and medical terminology based on 25 most common diagnosis names emerging in an electronic health records system. We also evaluated statistical properties of general terminology used in everyday life in respect to occurrence and importance to enable adaptive perspectives to medical knowledge. Our experiments exploit a conceptual co-occurrence network that we created based on a set of 93 medical texts about healthcare guidelines provided by The Finnish Medical Society Duodecim containing 57 679 unique conceptual links. We provide supplementing statistics of an extended range of Wikipedia articles and an n-gram analysis about the set of medical texts.

1 Introduction

We suggest that there is a strong need to develop intuitive computational methods to support online health queries by modeling patterns of creation, modification and retrieval of medical knowledge. We first give a brief overview about some features identified in online health queries, knowledge structures of Wikipedia online encyclopedia (https://www.wikipedia.org) and human knowledge processing. Then we report results of our experiments that evaluated properties of knowledge resources that can be used for building new semantically and behaviorally motivated resources of health guidance and clinical decision making. We measured some statistical properties of Wikipedia articles in respect to both general terminology and medical terminology. We also evaluated some statistical properties of general terminology used in everyday life and how those statistical properties can be exploited to enable adaptive perspectives to a conceptual co-occurrence network of medical knowledge that can address various needs of people who attempt to adopt and search suitable information for example by making online health queries. In addition we provide some supplementing statistics that we retrieved for an extended range of Wikipedia articles and an analysis of repeating sequences of words in the set of medical texts relying on n-grams.

2 Previous research

A survey in the USA population showed that 43.55 percent of people used the Internet to search for health information (Amante et al. 2015). A general search query length for personal computers has been found to be about 2.16-2.40 words and the number of search queries per session has been estimated to be about 2.52 for personal computers (Spink 2001). A general search query length for smart devices has been found to be about 2.3-2.35 words for cell phones and the number of search queries per session has been estimated to be about 1.6 for mobile devices (Kamvar & Baluja 2006). About two percent of online queries are health-related when this identification was carried out by matching with a large collection of medical terms (White & Horvitz 2009). When online queries and page visits about 12 common medical symptoms were investigated it was found that 3.6 percent of queries are health-related and 15.4 percent of visited pages are health-related (White & Horvitz 2009). Furthermore it turned out that 78.3 percent of all queries related to a medical symptom are typically made within a two weeks period since the first query is made for this symptom (White & Horvitz 2009).

The number of article entries in Wikipedia has surpassed all traditional encyclopedias, for example English edition of Wikipedia has 5.0 million articles as of January 2016 whereas the recent print editions of Encylopaedia Britannica contained about 65 000 articles (Berinstein 2006). In verifications the reliability of Wikipedia's factual contents has been considered to match traditional encyclopedias (Giles 2005; Chesney 2006). It has been shown that the content of Wikipedia follows approximately Zipf's law so that the exponent of probability density function $\alpha \approx 1.83$ and also Heaps' law so that the number of distinct words w(n) grows sublinearly with n (Serrano et al. 2009).

According to Zipf's law (Zipf 1935) pioneered by findings of Jean-Baptiste Estoup (Petruszewycz 1973), in large samples of natural language the frequency of any word f(z) is inversely proportional to its rank z based on the high-frequency list of all words, i.e. $f(z) \sim z^{-\zeta}$ with the scaling exponent ζ (Greek alphabet zeta) having a value of

about 1. When considering a word frequency distribution with a probability density function P(f) it appears in a form proportional to $f^{-\alpha}$ where the value of α (Greek alphabet alpha) has two variants: for universally shared words with $f > 10^{-5}$ there is $\alpha \approx 1 + 1/\zeta \approx 2$ whereas for significantly less frequently universally used words with $f < 10^{-5}$ there is $\alpha \approx 1.7$ (these values should hold for example in English language but some languages such as Chinese, Russian and Hebrew seem to have lower values) (Petersen et al. 2012). The behavior of Zipf's law has been explained by Simon (1955) with a model according to which a document is expanded either with a new word that has not yet occurred in the document with the probability of β (Greek alphabet beta) or with an old word with the probability of 1 - β, and this model is connected to the rank-frequency distribution of Zipf's law with the relation $\alpha = 1 + 1/(1 - \beta)$ (Simkin & Roychowdhury 2011).

According to Heaps' law (Heaps 1978), pioneered by Herdan's law (Herdan 1960), the number of distinct words in a document N_w is proportional to N_u^b, where N_u is the total number of words in a document and $b < 1$. When progressively excluding extremely rare words from a large document, the value of b increases from 0.5 to 1 and especially when having words with the frequencies of at least 1000 the value of b approaches 1 thus following the relation $b = 1 / \zeta$ that has been suggested to connect Zipf's law and Heaps' law (Petersen et al. 2012).

Human knowledge processing can be modeled with an adaptive associative network having a small-world structure and an average shortest path of 5.65-7.05 link steps between nodes (Morais et al. 2013). With an average reading speed of 200 words per minute (Lewandowski et al. 2003), the amount of knowledge discussed during a typical 20 minutes long consultation with a medical doctor can be coarsely estimated to correspond to reading a text containing about 4000 words. It has been identified that people respond more quickly to words that occur frequently in a language in respect to for example lexical decision, reading aloud, semantic categorization and picture naming (Duyck et al. 2008) and people recognize and produce more quickly words learned earlier in the life (Izura & Ellis 2002).

3 Method

We have carried out experiments to evaluate properties of knowledge resources that can be used for building new semantically and behaviorally motivated resources of health guidance and clinical decision making. Motivated by the findings of the previous research (Berinstein 2006; Giles 2005; Chesney 2006) it seems that Wikipedia online encyclopedia and especially its English edition are actively used by ordinary people for learning and checking facts about varied topics of life that also include medical knowledge that is at least partly applied for health guidance and clinical decision making. Therefore to evaluate properties of knowledge resources that can be used for building new semantically and behaviorally motivated resources of health guidance and clinical decision making we have decided to measure some statistical properties of encyclopedic articles of English edition of Wikipedia in respect to both general terminology and medical terminology. We then extend our analysis to evaluate some statistical properties of general terminology used in everyday life and how those statistical properties can be exploited to enable adaptive perspectives to medical knowledge and thus applied in the context of online health queries.

Now first we evaluate some statistical properties of Wikipedia articles in respect to general terminology. Relying on the title of the Wikipedia article, Table 1 shows the statistics of the highest-ranking Wikipedia articles about common nouns (or other sufficiently resembling groups of words that according to us can be considered as common nouns) from some alternative ranking-driven lists of Wikipedia articles based on empirical data that we have gathered and published originally in the publication (Lahti 2015a).

We have just evaluated some statistical properties of Wikipedia articles in respect to general terminology and now we next evaluate some statistical properties of Wikipedia articles in respect to medical terminology. A medical company Practice Fusion has reported 25 most common diagnosis names for the year 2011 emerging in its electronic health records system having a data set of over 7 million patients largely based on consultations with primary care physicians (Rowley 2011), the diagnosis names were listed in the order of decreasing frequency but the exact frequencies of the diagnosis names were not published. Table 2 shows these 25 most common diagnosis names in the order of decreasing frequency. Motivated by the previous results concerning the properties of online search queries and health queries discussed above we wanted to evaluate experimentally for Wikipedia articles corresponding to 25 most common diagnosis names some statistical properties reflecting patterns of creation, modification and retrieval of knowledge. We retrieved the statistics of Wikipedia articles about the diagnosis names as of January 2016 concerning page views per day (Wmflabs pageviews 2016), file size in bytes (Wmflabs xtools-articleinfo 2016), total number of edits (Wmflabs xtools-articleinfo 2016) and edits per year (Wmflabs xtools-articleinfo 2016). We coarsely estimated with a sample of Wikipedia articles that the file size shown for an article can possibly require about 4-10 bytes for each word that can be read in the article text.

Table 1. Some of the highest-ranking Wikipedia articles about common nouns in respect to the most viewed articles, the most edited articles, the longest articles in respect to file size and the most referenced articles in respect to receiving links from other articles (originally published in (Lahti 2015a)).

Most viewed articles in 2008 based on 210 analyzed days (Wikistics Falsikon 2009)		Most edited articles as of 30 July 2011 at 22:56 UTC (Wikipedia's pages with most revisions 2011)		Longest articles based on file size as of 29 July 2013 at 17:25 UTC (Wikipedia's long pages 2013)		Most referenced articles based on incoming internal links from articles (Wikipedia's most referenced articles 2011)	
article (ranking among all pages)	num-ber of views (page hits) per day	article (ranking among all pages)	number of edits (revisio ns)	article (ranking among all pages)	file size in bytes	article (ranking among all pages)	sum of direct links and links via redirects arriving from other articles
wiki (5)	140550	World War II (118)	21552	Plasmodium falciparum biology (9)	369920	geographic coordinate system (1)	662158
sex (17)	40141	Catholic Church (124)	21163	2000s (decade) (17)	325203	International Standard Book Number (3)	272923
2008 Summer Olympic Games (22)	28627	2006 Lebanon War (143)	19256	golden eagle (20)	314623	music genre (5)	191980
World War II (39)	21020	global warming (151)	18636	impalement (30)	304675	time zone (6)	190736
vagina (40)	20634	Jehovah's Witnesses (159)	17994	British literature (49)	280880	biological classification (7)	186918
penis (44)	19773	European Union (172)	17180	Iran-Iraq War (63)	268135	record label (9)	180716
masturbation (55)	18189	Islam (174)	17107	plug-in electric vehicle (68)	266102	animal (15)	138365
global warming (59)	17577	Christianit y (183)	16575	Gaza War (71)	265224	association football (17)	125106

Most viewed articles in 2008 based on 210 analyzed days (Wikistics Falsikon 2009)		Most edited articles as of 30 July 2011 at 22:56 UTC (Wikipedia's pages with most revisions 2011)		Longest articles based on file size as of 29 July 2013 at 17:25 UTC (Wikipedia's long pages 2013)		Most referenced articles based on incoming internal links from articles (Wikipedia's most referenced articles 2011)	
article (ranking among all pages)	num-ber of views (page hits) per day	article (ranking among all pages)	number of edits (revisio ns)	article (ranking among all pages)	file size in bytes	article (ranking among all pages)	sum of direct links and links via redirects arriving from other articles
anal sex (63)	17327	Hurricane Katrina (188)	16490	Euro zone crisis (79)	262361	binomial nomenclature (18)	124074
love (64)	17297	anarchism (204)	15905	sexuality in ancient Rome (80)	262267	record producer (20)	110761
sexual intercourse (65)	17190	Septembe r 11 attacks (207)	15851	Roman Empire (87)	261014	World War II (21)	109653
World War I (66)	17033	Iraq War (250)	14308	history of Western civilization (92)	258988	daylight saving time (22)	106392
Halloween (69)	16890	Scientolog y (253)	14261	War in Afghanistan (2001-present) (106)	254038	digital object identifier (27)	86406
pornography (79)	15776	Gaza War (256)	14221	Catholic Church and Nazi Germany (107)	253978	village (30)	77282
Olympic Games (80)	15751	World War I (267)	13988	Genie (feral child) (111)	252703	English language (31)	77087

In some cases the diagnosis name was redirected to another diagnosis name in Wikipedia and in some cases there was not a matching article for the diagnosis name but anyway a matching article for a partial or a partly different form of the diagnosis name. It turned out that the distribution of the statistical values of Wikipedia articles about diagnosis names have relatively varied patterns and we suggest that these patterns can usefully indicate some challenges that need to be solved when developing new adaptive methods that can support making online search queries and health queries. Already with this set of 25 most common diagnosis names it seems that there can be difficulties originating from having special terminology that affects how successfully online search queries and health queries can be carried out by ordinary people to find and understand relevant information that they need.

Based on the previous research (Morais et al. 2013; Duyck et al. 2008; Izura & Ellis 2002) it seems that statistical properties of general terminology used in everyday life and a personal language usage history and preference have an influence how the attention and associations of a human are directed and linked when he aims to search suitable information and to create and modify knowledge entities. Now we next evaluate some statistical properties of general terminology used in everyday life in respect to two properties that are occurrence and importance. Table 3 shows a sample of the highest-ranking common nouns we have gathered experimentally from students having ages in the range of 15-18 years (n=103) and published originally in the publication (Lahti 2014a). We asked each student to freely associatively write a list of 20 most important common nouns concerning the topic "life" (excluding the concept "life" itself) and then we asked everyone to review his own list and to give to each concept a ranking value representing a "measure of importance" ranging from 1 to 20 (value 1 meaning the most important). The students produced 621 unique nouns having altogether 1777 occurrences in word lists. We we will refer to these 621 unique nouns as "nouns of everyday life". Table 3 shows some of the highest-ranking common nouns in a decreasing order based on occurrences in word lists and the sum of measures of importance (a greater value meaning more important) that was computed based on inverse values so that the values in the range of 1-20 were translated to an inverse range of 21-1.

We have just evaluated some statistical properties of general terminology used in everyday life and now we next evaluate how those statistical properties can be exploited to enable adaptive perspectives to medical knowledge that can address various needs of people who attempt to search and adopt suitable information for example by making online health queries. Based on a set of 93 medical texts containing 85 055 words about healthcare guidelines given by Terveyskirjasto provided by The Finnish Medical Society Duodecim ("Käypä hoito, potilasversiot"; retrieved in January 2016 from http://www.terveyskirjasto.fi/terveyskirjasto/tk.koti?p_osio=109&p_teos=khp) we have created in our previous work (Lahti 2016b) a relatively comprehensive conceptual co-occurrence network about health-related topics containing 57 679 unique conceptual links. There were 2014 unique nouns having at least 3 occurrences in the set of medical texts and these 2014 unique nouns had altogether 101 024 co-occurrences in shared sentences that formed 28 840 unique concept pairs enabling bidirectionally the creation of 57 679 unique conceptual links.

Table 2. 25 most common diagnosis names for the year 2011 emerging in an electronic health records system of Practice Fusion having a data set of over 7 million patients largely based on consultations with primary care physicians (Rowley 2011) and statistics of Wikipedia articles about diagnosis names as of January 2016 concerning page views per day, file size in bytes, total number of edits and edits per year.

25 most common diagnosis names (Rowley 2011)		Statistics of Wikipedia articles about diagnosis names as of January 2016			
Ranking position	*Name (redirected name in Wikipedia)*	*Page views per day* (Wmflabs pageviews 2016)	*File size in bytes* (Wmflabs xtools-articleinfo 2016)	*Total number of edits* (Wmflabs xtools-articleinfo 2016)	*Edits per year* (Wmflabs xtools-articleinfo 2016)
1	Hypertension	4999	86450	4600	344
2	Hyperlipidemia	1390	20236	354	32.8
3	Diabetes (Diabetes mellitus)	6372	67895	7133	495.3
4	Back pain	871	43532	1441	110
5	Anxiety	3344	43926	3541	241.6
6	Obesity	2985	130233	7521	553.9
7	Allergic rhinitis	919	28042	532	44.2
8	Reflux esophagitis (Esophagitis)	32	4474	147	11.6
9	Respiratory problems	0 (Respiratory disease 640)	0 (Respiratory disease 12958)	0 (Respiratory disease 1057)	0 (Respiratory disease 88.1)
10	Hypothyroidism	4713	57643	2264	169.3
11	Visual refractive errors	0 (Refractive error 209)	0 (Refractive error 12558)	0 (Refractive error 184)	0 (Refractive error 14.2)
12	General medical exam	0 (Physical examination 529)	0 (Physical examination 17439)	0 (Physical examination 455)	0 (Physical examination 35)
13	Osteoarthritis	2425	70715	2100	177.1
14	Fibromyalgia/myositis, neuritis	0 (Fibromyalgia 6169; Myositis 348; Neuritis 185)	0 (Fibromyalgia 100744; Myositis 2100; Neuritis 2005)	0 (Fibromyalgia 4383; Myositis 79; Neuritis 58)	0 (Fibromyalgia 355.1; Myositis 8; Neuritis 5.8)
15	Malaise and fatigue	0 (Malaise 1387; Fatigue 134)	0 (Malaise 2022; Fatigue 799)	0 (Malaise 389; Fatigue 82)	0 (Malaise 34.7; Fatigue 6.9)
16	Pain in joint (Arthralgia)	1	8955	236	19.7
17	Acute laryngopharyngitis	0 (Upper respiratory tract infection 1225)	0 (Upper respiratory tract infection 15756)	0 (Upper respiratory tract infection 447)	0 (Upper respiratory tract infection 34.4)
18	Acute maxillary sinusitis (Sinusitis)	14	46988	1617	113.7
19	Major depressive disorder	3934	179596	9546	672.3
20	Acute bronchitis	678	14142	1119	90.4
21	Asthma	2978	94998	6755	488.6
22	Depressive disorders (Mood disorder)	14	51245	51245	49.7
23	Nail fungus (Onychomycosis)	12	21024	956	91.1
24	Coronary atherosclerosis (Atherosclerosis)	13	77092	1994	149.7
25	Urinary tract infection	4121	51988	1895	126.3

Table 3. Some of the highest-ranking common nouns of students (n=103) in a decreasing order based on occurrences in word lists and the sum of measures of importance (originally published in (Lahti 2014a)).

Mentioned concepts, occurrences in word lists of students (n=103)						Important concepts, the sum of measures of importance for students (n=103)					
concept	*frequency*	*concept*	*frequency*	*concept*	*frequency*	*concept*	*sum value*	*concept*	*sum value*	*concept*	*sum value*
family	53	sun	16	cat	10	family	903	child	202	sorrow	104
friend	49	dog	15	air	9	friend	821	joy	195	learning	103
work	41	hobby	15	clock	9	love	525	hobby	188	book	99
death	40	house	15	learning	9	work	445	study	186	computer	99
love	33	education	14	mother	9	water	408	happiness	179	clock	98
school	33	health	14	summer	9	food	396	education	172	cloth	95
food	31	money	14	television	9	death	363	house	147	free_time	91
water	31	sorrow	14	living	8	school	362	plant	136	holiday	91
animal	29	study	14	music	8	human	335	mother	133	music	91
human	24	computer	13	party	8	birth	321	money	130	party	87
birth	23	plant	12	religion	8	nature	303	air	121	emotion	86
nature	21	car	11	city	7	animal	285	dog	118	fun	85
home	18	happiness	11	cloth	7	home	237	world	106	summer	85
child	16	tree	11	elderness	7	health	225	father	105	tree	85
joy	16	book	10	environment	7	sun	224	living	105	purpose	84

It turned out that among 2014 unique nouns of the conceptual co-occurrence network there were 161 unique nouns that belonged to the set of 621 unique “nouns of everyday life” gathered from the students (n=103). In the conceptual co-occurrence network of 57 679 unique conceptual links there appeared to be 1994 unique links that traversed between nouns that belonged to the set of 621 unique “nouns of everyday life”. Table 4 shows some of the highest-ranking conceptual links of 1994 unique links in a decreasing order based on occurrences in word lists and the sum of measures of importance gained from the students for the nouns (links going to the opposite directions are shown combined bidirectionally). For each link we computed the sum of values of two nouns forming the link in respect to occurrences in word lists and the sum of measures of importance and then we sorted the links in a descending order based on the sum of values of two nouns forming the link.

Table 4. Some of the highest-ranking conceptual links of 1994 unique links of the conceptual co-occurrence network based on the set of 93 medical texts about healthcare guidelines in a decreasing order based on occurrences in word lists and the sum of measures of importance gained from the students (n=103) for the nouns. Links going to the opposite directions are shown combined bidirectionally.

Conceptual links, occurrences in word lists of students (n=103)						Conceptual links, the sum of measures of importance for students (n=103)					
link	*sum*	*link*	*sum*	*link*	*sum*	*link*	*sum*	*link*	*sum*	*link*	*sum*
child↔family	69	family↔weekday	54	material↔work	43	child↔family	1105	food↔health	621	disease↔work	473
family↔ young_(person)	58	home↔school	51	possibility↔work	43	drink↔family	938	education↔work	617	care↔work	465
drink↔family	56	light↔work	48	stress↔work	43	family↔relationship	938	home↔school	599	difficulty↔work	465
family↔problem	56	child↔food	47	activity↔work	42	family↔year	933	child↔food	598	eating↔food	465
family↔relationship	56	disease↔work	47	beginning↔work	42	family↔ young_(person)	933	mother↔water	541	learning↔school	465
family↔year	56	death↔heart	46	care↔work	42	care↔family	923	food↔mother	529	reason↔work	465
education↔work	55	death↔purpose	45	death↔possibility	42	family↔problem	919	birth↔child	523	doctor_(physician)↔ work	464
family↔information	55	food↔health	45	doctor_(physician)↔ work	42	family↔information	917	light↔work	512	survival↔work	462
health↔work	55	goal_(to_achieve)↔w ork	45	learning↔school	42	family↔habit	915	goal_(to_achieve)↔ work	509	activity↔work	461
care↔family	54	death↔hospital	44	longness↔work	42	family↔item	915	cloth↔food	491	beginning↔work	460
family↔habit	54	death↔ physical_training	44	need↔work	42	family↔need	912	work↔ working_place	490	stress↔work	460
family↔item	54	spirit↔work	44	reason↔work	42	family↔space	912	environment↔water	483	thought↔work	455
family↔need	54	work↔working_place	44	regeneration↔work	42	family↔weekday	912	possibility↔work	482	material↔work	454
family↔space	54	work↔year	44	survival↔work	42	family↔thing	904	spirit↔work	480	need↔work	454
family↔thing	54	difficulty↔work	43	thought↔work	42	health↔work	670	work↔year	475	food↔time	451

4 Discussion and future work

Our evaluations indicate that when aiming to develop computational methods that enable generating adaptive perspectives to medical knowledge it can be useful to analyze statistical properties dealing with various knowledge resources such as Wikipedia articles in respect to both general terminology and medical terminology as well as general terminology used in everyday life and medical texts offering reliable healthcare guidelines. Naturally, the approach we have used for the analysis and our results should be seen as an illustration about one of many diverse possibilities for developing adaptive perspectives to medical knowledge that can hopefully be explored more thoroughly in future research.

The statistics just discussed for 25 most common diagnosis names can be contrasted with some supplementing statistics that we retrieved for a broader range of Wikipedia articles. As of January 2016 the domain wikipedia.org is the 7th most popular web site globally according to Alexa Internet web traffic report (Alexa Internet 2016) and the domain wikipedia.org has an estimated traffic of 7.3 billion visits per month according to Rank2traffic web traffic report (Rank2traffic 2016). According to Wmflabs wikitrends (2016) on English edition of Wikipedia in the year 2015 the most viewed article was "Deaths in 2015" having 20 686 307 views per year and the second most viewed article was "Facebook" having 10 769 172 views per year. According to Wmflabs topviews (2016) on English edition of Wikipedia in January 2016 the top 100 most viewed articles had each a minimum of 1 083 646 views per month, the top 1000 articles had each a minimum of 177 871 views per month, the top 2000 articles had each a minimum of 61 791 views per month and the top 3000 articles had each a minimum of 31 350 views per month.

For the conceptual co-occurrence network based on the set of 93 medical texts about healthcare guidelines we created perspectives emphasizing occurrences in word lists and the sum of measures of importance gained from

the students for the nouns and these results can be contrasted with some supplementing analysis relying on n-grams that are repeating sequences of words in a text. The set of 93 medical texts contained 85 055 words and 2014 unique nouns having at least 3 occurrences. Based on the set of 93 medical texts we created a simplified "noun text" so that we excluded all other concepts than 2014 unique nouns and end-of-sentence tokens (this approach is introduced in (Lahti 2016b)). Thus we gained a "noun text" consisting of a sequence of 45 241 nouns of which 12 141 were end-of-sentence tokens. Table 5 shows some of the highest-ranking n-grams we have identified in the "noun text" based on the set of 93 medical texts including unigrams, bigrams, trigrams, four-grams and five-grams (we excluded some n-grams concerning the publication format of the medical texts).

Table 5. Some of the highest-ranking n-grams in a "noun text" based on the set of 93 medical texts about healthcare guidelines. Frequencies are shown for unigrams, bigrams, trigrams, four-grams and five-grams.

Unigrams	*Freq.*	*Bigrams*	*Freq.*	*Trigrams*	*Freq.*
care	1326	care¤base	61	care¤base¤medicine	23
patient	685	child¤young	51	lower_limb¤vein¤hypofunction	11
symptom	567	patient¤symptom	34	calcium¤d_vitamin¤intake	6
child	389	type¤diabetes	33	pregnancy¤duration¤week	5
time	275	smoking¤quitting	31	pain¤permission¤limit	5
medicine	262	care¤aim	29	child¤young¤care	5
year	244	pregnancy¤time	29	tobacco_addiction¤tobacco¤quitting	5
trace	237	year¤time	28	smoking¤quitting¤symptom_of_quitting	4
disease	232	condition¤situation	27	blood¤coagulation¤clot	4
example	205	patient¤care	25	patient¤position¤right	4
pain	201	part¤patient	24	symptom¤examination¤care	4
reason	199	number¤light	24	wasp¤bee¤sting	4
surgery	198	base¤medicine	24	man¤percent¤women	4
month	198	year¤age	23	hair_follicle¤mouth¤inflammation	4
examination	192	down¤syndrome	22	size¤life¤time	4
supplement	190	type¤diabetic	22	child¤young¤obesity	4
medication	181	examination¤care	21	care¤surgery¤surgery	4
tooth	180	surgery¤trace	21	knee¤knee¤hip_osteoarthritis	4
use	179	mouth¤period	21	hand¤forearm¤epicondylitis	4
doctor	171	care¤patient	20	vagina¤period¤function	4
risk	169	acute¤kidney_failure	20	half¤year¤time	4
medication	168	care¤medicine	20	use¤pregnancy¤time	4
need	165	week¤time	20	pregnancy¤breastfeeding¤time	4

Four-grams	*Freq.*
physical_activity¤osteoarthritis¤care¤base	3
exercise¤physical_activity¤osteoarthritis¤care	3
virus _infection¤cough¤acute¤bronchitis	3
child¤first¤expiratory_dyspnea¤bronchiolitis	3
year¤child¤first¤expiratory_dyspnea	3
rehabilitation¤mobility_exercise¤surgery¤day	3
adhd¤activity¤attention¤disorder	3
nicotine_addiction¤quitting_treatment¤ tobacco_addiction¤tobacco	3
quitting_treatment¤tobacco_addiction¤ tobacco¤quitting	3
obesity¤adult¤obesity¤adult	3
pregnancy¤duration¤week¤abortion	3
blood¤coagulation¤clot¤medicine	3
myocardial_infarction¤stroke¤year¤time	3
mouth_cancer¤risk_factor¤tobacco¤alcohol	3
law¤patient¤position¤right	3
Five-grams	*Freq.*
exercise¤physical_activity¤osteoarthritis¤care¤ base	3
nicotine_addiction¤quitting_treatment¤ tobacco_addiction¤tobacco¤quitting	3
year¤child¤first¤expiratory_dyspnea¤bronchiolitis	3

We have provided experimental results about properties of knowledge resources that can be used for building new semantically and behaviorally motivated resources of health guidance and clinical decision making. Due to space constraints this article can show only a part of the results of our semantic analysis and thus we have decided to publish more extensive listings of our results that are available as open data in a separate publication (Lahti 2016c). Based on the previous research and our results we suggest that exploration, adoption, exploitation and further development of knowledge should be supported with computational methods that help a person to relate new knowledge to his previous knowledge and to address conceptual topics that are frequently present in his everyday life and important for him. We suggest that future research should emphasize multidisciplinary development of modular adaptive computational methods for personalized healthcare and health informatics relying on diverse data resources of medical work that enable efficient and flexible experimental testing, distribution, modification and application by both scientific community and ordinary people.

References

Alexa Internet (2016). Web traffic report for January 2016. http://www.alexa.com/siteinfo/wikipedia.org.

Amante, D., Hogan, T., Pagoto, S., English, T., & Lapane, K. (2015). Access to care and use of the Internet to search for health information: Results from the US National Health Interview Survey. Journal of Medical Internet Research, 17(4): e106. doi:10.2196/jmir.4126. http://www.ncbi.nlm.nih.gov/pmc/articles/PMC4430679/

Berinstein, P. (2006). Wikipedia and Britannica - The kid's all right (and so's the old man). Searcher, 14(3). Information Today, Inc. http://www.infotoday.com/searcher/mar06/berinstein.shtml

Chesney, T. (2006). An empirical examination of Wikipedia's credibility. First Monday, 11(11).

Duyck, W., Vanderelst, D., Desmet, T., & Hartsuiker, R. (2008). The frequency effect in second-language visual word recognition. Psychonomic Bulletin & Review, 15(4), 850-855. http://users.ugent.be/~wduyck/articles/DuyckVanderelstDesmetHartsuiker2008.pdf

Giles, G. (2005). Internet encyclopaedias go head to head. Nature, 438, 7070, 900-901.

Heaps, H. (1978). Information retrieval: computational and theoretical aspects. Academic Press, New York, USA.

Herdan. G. (1960). Type-token mathematics. Mouton, The Hague, the Netherlands.
Izura, C., & Ellis, A. (2002). Age of acquisition effects in word recognition and production in first and second languages. Psicológica, 23, 245-281. http://www.uv.es/revispsi/articulos2.02/4.IZURA%26ELLIS.pdf
Kamvar, M., & Baluja, S. (2006). A large scale study of wireless search behavior: Google mobile search. Proc. Conference on Human Factors in Computing Systems (CHI '06), 701–709. http://www1.cs.columbia.edu/~mkamvar/publications/CHI_06.pdf
Lahti, Lauri (2014a). Educational exploration based on conceptual networks generated by students and Wikipedia linkage. Proc. World Conference on Educational Multimedia, Hypermedia and Telecommunications 2014 (EdMedia 2014). 23–27 June 2014, Tampere, Finland (eds. Herrington, J. et al.), 964–974. ISBN 978-1-939797-08-7. Association for the Advancement of Computing in Education (AACE), Chesapeake, VA, USA. http://www.editlib.org/p/147608/; http://urn.fi/URN:ISBN:978-1-939797-08-7
Lahti, Lauri (2015a). Computer-assisted learning based on cumulative vocabularies, conceptual networks and Wikipedia linkage. Doctoral dissertation. Department of Computer Science, Aalto University School of Science, Finland. Unigrafia Oy, Helsinki, Finland. ISBN 978-952-60-6163-4 (printed), ISBN 978-952-60-6164-1 (pdf). http://urn.fi/URN:ISBN:978-952-60-6164-1
Lahti, Lauri (2016b). Semantic modeling of healthcare guidelines to support health literacy and patient engagement. Proc. Global Learn 2016: Global Conference on Learning and Technology, 28-29 April 2016, Limerick, Ireland. Association for the Advancement of Computing in Education (AACE), Chesapeake, VA, USA. https://www.learntechlib.org/p/172737/; http://urn.fi/URN:NBN:fi:aalto-201603291477
Lauri, Lahti (2016c). Supplement to Lauri Lahti's conference article ”Supporting online health queries by modeling patterns of creation, modification and retrieval of medical knowledge”, to appear online at http://aaltodoc.aalto.fi.
Lewandowski, L., Codding, R., Kleinmann, A., & Tucker, K. (2003). Assessment of reading rate in postsecondary students. Journal of Psychoeducational Assessment, 21, 134-144.
http://www.iapsych.com/wj3ewok/LinkedDocuments/lewandowski2003.pdf
Morais, A., Olsson, H., & Schooler, L. (2013). Mapping the structure of semantic memory. Cognitive Science, 37, 125-145.
Petruszewycz, M. (1973). L'histoire de la loi d'Estoup-Zipf: documents. Mathématiques et sciences humaines, 44, 41-56.
Petersen, A., Tenenbaum, J., Havlin, S., Stanley, H., & Perc, M. (2012). Languages cool as they expand: allometric scaling and the decreasing need for new words. Scientific Reports 2, 943.
http://www.matjazperc.com/publications/ScientificReports_2_943.pdf
Rank2traffic (2016). Web traffic report for January 2016. http://www.rank2traffic.com/wikipedia.org
Rowley, R. (2011). The 25 most common diagnoses. Robert Rowley, Chief Medical Officer, Practice Fusion EMR. Practice Fusion Blog, posted on 9 February 2011. http://www.practicefusion.com/blog/25-most-common-diagnoses/
Serrano, M., Flammini, A., & Menczer, F. (2009). Modeling statistical properties of written text. Public Library of Science ONE (PLoS ONE), 4(4): e5372. http://www.plosone.org/article/info%3Adoi%2F10.1371%2Fjournal.pone.0005372
Simkin, M., & Roychowdhury, V. (2011). Re-inventing Willis. Physics Reports 502, 1-35.
http://arxiv.org/ftp/physics/papers/0601/0601192.pdf
Simon, H. (1955). On a class of skew distribution functions. Biometrika, 42, 425.
Spink, A., Wolfram, D., Jansen, M., & Saracevic, T. (2001). Searching the web: The public and their queries. Journal of the American Society for Information Science and Technology 52 (3): 226–234.
https://faculty.ist.psu.edu/jjansen/academic/jansen_public_queries.pdf
White, R., & Horvitz, E. (2009). Cyberchondria: Studies of the escalation of medical concerns in Web search. Proc. ACM Transactions on Information Systems (TOIS), 27(4), article 23. http://research.microsoft.com/en-us/um/people/ryenw/papers/whitetr-2008-178.pdf
Wikipedia’s long pages (2013). Long pages based on file size, generated live for Wikipedia’s special page Special:LongPages as of 29 July 2013 at 17:25 UTC. Online available at https://en.wikipedia.org/w/index.php?title=Special:LongPages
Wikipedia’s the most referenced articles (2011). The most referenced articles based on incoming internal links from articles, relying on sum of direct links and links via redirects. Generated live for Wikipedia’s special page Wikipedia:Most_Referenced_Articles as of 21 August 2011. Online available at http://en.wikipedia.org/wiki/Wikipedia:Most_Referenced_Articles
Wikipedia’s pages with the most revisions (2011). Pages with the most revisions, limited to the first 1000 entries. Generated live for Wikipedia’s special page Wikipedia:Database_reports/Pages_with_the_most_revisions as of 30 July 2011 22:56 UTC. Online available at http://en.wikipedia.org/wiki/Wikipedia:Database_reports/Pages_with_the_most_revisions
Wikistics Falsikon (2009). Page hits per day for en.wikipedia in year 2008. Based on 210 analysed days, requests counted by Squid servers. Online available at http://wikistics.falsikon.de/2008/wikipedia/en/.
Wmflabs pageviews (2016). Wikimedia Tool Labs Pageviews Analysis. https://tools.wmflabs.org/pageviews/#project=en.wikipedia.org&platform=all-access&agent=user&start=2016-01-01&end=2016-01-31&pages=Hypertension
Wmflabs topviews (2016). Wikimedia Tool Labs Topviews Analysis. http://tools.wmflabs.org/topviews/#project=en.wikipedia.org&platform=all-access&start=2016-01-01&end=2016-01-31&excludes=
Wmflabs wikitrends (2016). Wikimedia Tool Labs Wikitrends. Most viewed articles on English Wikipedia 2015. http://tools.wmflabs.org/wikitrends/2015.html
Wmflabs xtools-articleinfo (2016). Wikimedia Tool Labs Article Info (page history). https://tools.wmflabs.org/xtools-articleinfo/?article=Hypertension&project=en.wikipedia.org
Zipf, G. (1935). The psychobiology of language: an introduction to dynamic philology. Houghton-Mifflin, Boston, Massachusetts, USA.