Lahti, Lauri (2016a). Evaluation of semantic dependencies in a conceptual co-occurrence network of a medical vocabulary. Proc. 5th International Conference on Human Computing, Education and Information Management System (ICHCEIMS 2016), 27-28 March 2016, Sydney, Australia. Open access in Aaltodoc publication archive at http://aaltodoc.aalto.fi.

Evaluation of semantic dependencies in a conceptual co-occurrence network of a medical vocabulary

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

Abstract*:*

The amount of medical knowledge is constantly growing thus providing new hope for people having health-related problems. However a challenge is to develop flexible methods to facilitate managing and interpreting large medical knowledge entities. There is a need to enhance health literacy by developing personalized health support tools. Furthermore there is a need to assist decision-making with decision support tools. The recent and on-going changes in everyday life both on technological and societal levels (for example adoption of smart phones and personal mobile medical tracking devices, social networking, open source and open data initiatives, fast growth of accumulated medical data, need for new self-care solutions for aging European population) motivate to invest in the development of new computerized personalized methods for knowledge management of medical data for diagnosis and treatment. To enable creation of new adaptive personalized health support tools we have carried out an evaluation of semantic dependencies in a conceptual co-occurrence network covering a set of concepts of a medical vocabulary with experimental results ranging up to 2994 unique nouns, 82814 unique conceptual links and 200000 traversed link steps.

Keywords–personalized healthcare; health informatics; patient guidance; conceptual network; the shortest path

I. INTRODUCTION

The amount and complexity of gathered and analyzed biomedical data is growing fast since according to estimates the doubling time of medical knowledge was 7 years in 1980 and 3,5 years in 2010 and will be 73 days in 2020 (Densen 2011). There is a need for developing new computational analysis methods to address challenges of managing and analyzing increasing knowledge resources effectively. A recent trend in biomedical research has been emphasizing integrative and translational research which deals with management and analysis of large-scale heterogeneous data sets but many frameworks suggested for this approach are still relatively experimental and needing increased validation efforts (Sulakhe et al. 2014). It has been considered that the current clinical research system is not able to provide sufficiently efficiently high-quality evidence to support decision making about health care and this has motivated initiatives to introduce more patient-centered research networks and strategies (Califf 2014).

A knowledge engineering process has been considered as a useful theoretical frame for biomedical research since this process aims to gather suitable knowledge, to represent this knowledge in a computable form, to implement knowledge-based agents for this knowledge and to validate the output of the agents against some reference standards (Payne 2012). The knowledge engineering process can manage various types of knowledge including three typical categories: conceptual knowledge, procedural knowledge and strategic knowledge. In the context of a support system for clinical decision making the conceptual knowledge might cover symptoms and

diagnoses and relationships between them, the procedural knowledge might cover algorithms to process the just mentioned knowledge and the strategic knowledge might cover logic that enables practically applying the just mentioned knowledge (Payne 2012).

The policy brief of World Health Organization (WHO) argues that about 30 percent of Europeans have a long-standing illness or health problem and there is a need to support health literacy by developing personalized health information that is appropriate, timely, relevant and reliable, to support participation in decision-making by developing communication methods of medical professionals and developing decision support tools for the patients to empower their needs in consultation with a medical professional, and to support educating self-management skills developing methods for peer-support groups and changing people's health behavior that can benefit from social interactive computer-based applications (Coulter et al. 2008).

II. PREVIOUS RESEARCH

A. Modeling semantic dependencies in health information

Case-based reasoning systems have been considered as a useful approach to support health sciences domains (Marling et al. 2014). In automated pattern recognition of medical imaging some promising results have been gained with models that support interactive search when a previously computed index is not available (Goode et al. 2008). To manage imprecise, uncertain and incomplete medical data a promising approach has been application of rough set theory that can represent dependencies of data without requiring much supplementing information (Tripathy et al. 2011). In drug safety research that requires dealing with hundreds of ontologies at the same time and expanding query inputs and search space it has been noted that some promising results can be gained with models relying on theories of formal concept analysis methods and semantic query expansion (Curé et al. 2015).

Semantic web technologies have been successfully applied in the domains of neuroscience and biomedicine but challenges exist concerning for example limited standardization and coverage of user base (Ruttenberg et al. 2007). Various alternative and complementing perspectives have been identified to model decision making in the context of neuroscience (Kalencher & Tobler 2008). N-grams that aim to represent occurrences of similar consecutive patterns in texts have been successfully used to classify medical texts about chief complaints of patients based on groupings of international classification of diseases (Brown et al. 2010). N-grams have been also used to classify heart diseases so that heart beat signals have been first converted to symbolic representation with k-means clustering algorithm and then symbolic sequences based on n-grams have been categorized (Huang et al. 2012).

Methods for keyphrase extraction have been proposed to help text summarization, automatic indexing, clustering and classification (Sarkar 2013). In keyphrase extraction it is a challenge to identify descriptive expressions and some proposed solutions include using for example part-of-speech taggers, parsers, naive Bayes classifiers, decision trees, the position in a document, the presence in ontologies or in Wikipedia texts and especially a combination of various features has been successful (Chuang et al. 2012). An experiment about keyphrase extraction noted that people most commonly described documents with bigrams but the use of

unigrams increased along the number and diversity of documents and that for text visualization commonly used raw term frequency and tf-idf measures could be outperformed with for example a G2 measure (measuring the significance of the document term in respect to a reference corpus) or a linear combination of log term frequency and Web commonness (Chuang et al. 2012).

Scale-free small-world networks appear to emerge inherently in many natural processes and social phenomena, such as social networking. In small-world networks the average distance between nodes is small and in scale-free networks the average distance can become especially small (Newman 2000; Cohen & Havlin 2003). In previous research the small-world networks have been suggested to have an important role as organizing and processing knowledge in biological neural networks (Pajevic & Plenz 2009; Stratton & Wiles 2010; Wang et al. 2010). To support medical diagnosis and treatment there have been made alternative network-based modelings for example in respect to collaboration between doctors or connections between human symptoms and diseases (O'Malley & Marsden 2008; Zhou et al. 2014).

It has been shown that co-occurrence of concepts in sentences can be represented with networks that have scale-free small-world properties and in which the average distance between any two concepts is about 2-3 link steps (Ferrer i Cancho & Solé 2001). It has been identified that semantic network models relying on subject-verb-object form generated based on a medical publication database can be successfully used to represent semantic relatedness reflecting human judgment and these models can in certain cases outperform three popular other models using path-based, statistical or context vector methods (Workman et al. 2013). Based on a set of collective discourse examples representing various individual perspectives it was shown that the network of co-occurring perspectives followed small-world properties and had a high clustering of different perspectives (Qazvinian & Radev 2012). It has been experimentally shown that a method based on network analysis of co-occurring terms and their ranking can effectively classify biomedical articles in large databases thus supporting making queries about genes, chemicals or diseases (Hsu & Kao 2013). Cumulatively layered models applicable to analyze similarities in various biomedical network structures have been developed with kernel and convolution methods that can rely on subgraph, subtree and shortest paths approaches as well as using neural language models, continuous bag-of-words models, skip-gram models and deep graph kernels (Yanardag & Vishwanathan 2015).

Linguistic data can be successfully analyzed with various kinds of multidimensional models relying on semantic spaces (Mikolov et al. 2013; Karlgren et al. 2014). Evaluations have been carried out to explore how distributional semantics can be used to facilitate natural language processing concerning electronic health records in respect to synonym extraction of medical terms, assignment of diagnosis codes and identification of adverse drug reactions (Henriksson 2013). An important modeling task in the medical domain has been to develop automated methods for diagnostic coding (Stanfill et al. 2010). Analysis relying on clinical narratives has been considered as a promising source for classifying medical data for example with methods of k-nearest-neighbor, relevance feedback and Bayesian independence (Larkey & Croft 1996).

Models relying on narrative networks can represent medical organizational workflow processes on social and technological level relying on collections of stories to better identify structures and relationship patterns from various points of view (including patients, doctors, nurses etc.) (Hayes et al. 2011). The models of narrative

networks are related to for example structuration theory (Giddens 1984), actor-network theory (Latour 2005), the theory of organizational routines (Feldman & Pentland 2003), valued directed graphs (Ablell 2004) and first-order Markov models (Abbott 1992). In the formation of narrative networks various levels of granularity can be observed built modularily based on actants, narrative fragments and narratives, and four common building phases include choosing a focal phenomenon with boundaries, choosing a point of view, collecting narratives with code fragments and relating the nodes of narratives by a sequence (Hayes et al. 2011).

It has been suggested that seven guidelines for designing live routines (Pentland & Feldman 2008) can be usefully implemented with the methods of narrative networks (Hayes et al. 2011) thus enabling the following activities in the context of medical workflows and when designing a medical records system: to describe generalized abstract functions of practice and technological use, to consider each individual viewpoint, to consider relationships between observed actions and abstract patterns, to examine pathways as important sites of intervention or reinforcement, to explore decision points as probable locations of innovative problem solving and using autonomity, to identify points or narrative fragments where alternative pathways can be harmful for the organization, and to enable all participating people to maintain accommodating to change and new routines.

B. Properties of online search queries and health queries

It has been identified that online health information-seeking is more frequent for women and people who are more educated, have higher income and have high-speed access to Internet at home and work (Atkinson et al. 2009; Higgins et al. 2011; Wangberg et al. 2008; Kummervold et al. 2008). Besides health information-seeking made by a person for himself, a half of the searches is carried out on behalf of a friend or a relative (Sadasivam et al. 2013). A survey in the USA population showed that 43.55 percent of people used the Internet to search for health information and 3.63 percent used online health chat rooms to learn about health topics (Amante et al. 2015). In the age groups 18-35 years and 35-60 years over 50 percent reported using Internet searches or health chat rooms whereas in the age group 60 years or older 31.35 percent reported this behavior (Amante et al. 2015). People who reported experiencing a delay in getting an appointment soon enough, being not accepted as a new patient or not having acceptance for the insurance or having the doctor's office closed when access would be needed had over two times the probability to use the Internet to search for health information than people who did not report these challenges (Amante et al. 2015).

A general search query length for personal computers has been identified to be on average 2.16-2.40 words (Spink 2001) and for smart devices about 2.3-2.35 words for cell phones (Kamvar & Baluja 2006). The number of search queries per session has been estimated to be about 2.52 for personal computers (Spink 2001) and about 1.6 for mobile devices (Kamvar & Baluja 2006). For personal computers 51.8 percent of search queries were unique queries (Spink 2001) and among modified search queries in 41.6 percent the number of terms was increased and in 25.9 percent was decreased and in 32.5 percent remained the same (Spink 2001). An average health query length was for personal computers 2.90-4.82 words and for smart devices 3.29-5.33 words (Ashutosh et al. 2014). Health queries made from a smart device are longer and more descriptive than queries made from a personal computer, more queries are made from smart devices than from personal computers, and queries are made by women and children more often than men or other age groups (Ashutosh et al. 2014). Health

queries of question-related words were more frequent for smart devices that personal computers (Ashutosh et al. 2014). In queries of question-related words the expressions when and can were more frequent for smart devices than personal computers whereas the expressions what, is and does were more frequent for personal computers than smart devices (Ashutosh et al. 2014). In health queries without repetitions 31 percent included at least one spelling mistake, 37-47 percent at least one verb, and 45.66-48.50 percent at least one adjective (Ashutosh et al. 2014).

It was found that about two percent of online queries are health-related when based on matching with a large collection of medical terms (White & Horvitz 2009). When analyzing online queries and page visits concerning 12 common medical symptoms it was identified that 3.6 percent of queries are health-related and 15.4 percent of visited pages are health-related, and that 78.3 percent of all queries related to a medical symptom are typically carried out within a two weeks period since the initial query is made for this symptom (White & Horvitz 2009). 5.3 percent of symptom queries led to an uncommon and serious explanation of medical condition (i.e. escalation), 7.4 percent led to a high-likelihood and non-serious explanation of medical condition (i.e. non-escalation) and 87.3 percent led to a no change (White & Horvitz 2009). Among people making the searches 32.9 percent made escalating symptom queries and 70.1 percent made non-escalating symptom queries (with an overlap of 3 percent) (White & Horvitz 2009).

A health query session that contained at least one escalation had on average a duration of 3801 seconds, 24.8 query iterations, 29.2 visited pages and among the visited pages 39.1 percent were medical pages and 25.1 percent medical pages from trusted sources (White & Horvitz 2009). A health query session that contained at least one non-escalation had on average a duration of 3412 seconds, 16.6 query iterations, 16.1 visited pages and among the visited pages 39.2 percent were medical pages and 19.1 percent medical pages from trusted sources (White & Horvitz 2009). During a health query session the distance from a symptom query to the first escalation was on average 2.3 queries, 2.2 page views or 132.7 seconds and the distance from a symptom query to the first non-escalation was on average 1.2 queries, 1.1 page views or 93.3 seconds (White & Horvitz 2009). In a personal medical search history it was measured that the searches about symptoms have on average a distance of 18.9 days or 22.8 search sessions, the searches about serious illness have on average a distance of 19.0 days or 20.5 search sessions and the searches about common explanations have on average a distance of 11.4 days or 12.6 search sessions (White & Horvitz 2009).

III. METHOD

In our previous work (Lahti 2015c) we have proposed a new computational method to support learning that relies on adaptive exploration of the shortest paths in conceptual networks that have been formed based on co-occurrences of concepts in a set of suitable text samples and selecting concepts corresponding to desired language ability levels. Our previous results were based on an experimentally generated network. For each of 3018 highest-ranking nouns of British National Corpus (BNC) available from lemmatized word lists of an online database (Leech et al. 2001) we queried an online database for Google Web 1T 5-gram database (FAU Erlangen-Nürnberg 2015) to identify all other nouns belonging to the same set of 3018 nouns that co-occur at the distance of at the most four words left or right, with the association measure of t-score and considering at most 50 highest-ranking

nouns having a frequency of at least 40. When identifying nouns we relied on just matching spelling and thus some non-nouns may have possibly become unintentionally considered as nouns and some nouns may have become unintentionally excluded. For all 3018 unique nouns we gained together a set of 54610 unique pairs of nouns (i.e. each pair corresponding to a co-occurrence of two nouns, both belonging to the set of 3018 unique nouns), and it appeared that 2994 of 3018 unique nouns occurred in these unique pairs of nouns.

Among 54 610 unique pairs of nouns for 26406 pairs of nouns (about 48 percent) there was a co-occurrence in both directions, thus 13203 connections were bidirectional. To enable a full bidirectional reach for all interconnecting link paths in the network we generated additional 28 204 unique pairs of nouns in the opposite direction for those 28204 unique pairs of nouns not originally having a co-occurrence in the opposite direction. Thus we carried out our further analysis based on a conceptual co-occurrence network containing altogether 82814 unique pairs of nouns (i.e. 82814 unique links) between 2994 unique nouns, and when considering each link as bidirectional there are 82814/2=41407 unique bidirectional pairs of nouns.

Motivated by our previous work (Lahti 2015c) that evaluated a conceptual co-occurrence network between a set of concepts belonging to a general vocabulary (British National Corpus) we wanted now to evaluate a conceptual co-occurrence network between a set of concepts belonging to a medical vocabulary. Wang et al. (2008) have defined a Medical Academic Word List (MAWL) based on the most frequently used medical academic words in a set of medical research articles containing 1093011 running words gathered from online resources. The Medical Academic Word List includes 623 word families that cover 12.24 percent of the part-of-speech tokens belonging to the set of research articles. In our current work we analyzed how the concepts of the Medical Academic Word List (Wang et al. 2008) are represented and connected in the conceptual co-occurrence network generated in our previous work (Lahti 2015c) containing 82814 unique pairs of nouns. It turned out that 257 of 623 unique concepts of the Medical Academic Word List occurred among 3018 highest-ranking nouns of British National Corpus (Leech et al. 2001) and thus also among the concepts of the conceptual co-occurrence network. For each of these 257 medically motivated concepts we wanted to identify the connectivity to all other concepts belonging to the same set of 257 concepts inside the conceptual co-occurrence network and therefore we defined a set of (257*256)/2=32896 concept pairs that we call as potentially explorable concept pairs.

Table 1 shows the highest-ranking 3018 nouns in BNC and the most frequent word families (concepts) of MAWL as well as the concepts of MAWL among 3018 nouns of BNC having the highest ranking among the nouns of BNC and among the concepts of MAWL. We provided each of 2994 unique nouns of the conceptual co-occurrence network with a ranking value corresponding to its ranking position among the highest-ranking 3018 nouns in BNC (given as consecutive ranking values in the range of 0-2993), called as a rank position in a general vocabulary. For each of 32896 potentially explorable concept pairs we calculated a summing ranking value that is a sum of the values of rank position in a general vocabulary for both concepts belonging to this concept pair. For further analysis we extracted those potentially explorable concept pairs that had a summing ranking value of at most 342 and thus we gained a set of 1005 unique concept pairs. Ten concept pairs having the lowest summing ranking values (and thus having the highest ranking positions) were: approach@area (78), acid@area (81), cell@period (90), clinic@period (94), cell@community (97), factor@period (97), method@period (98), clinic@community (101), dose@period (101) and gene@period (102). Please note that in the conceptual co-

occurrence network the connectivity is expected be enabled in both direction (i.e. both from conceptA to conceptB and from conceptB to conceptA) and in the notation for a concept pair conceptA@conceptB the order of the concepts is based on an alphabetic order. In the set of 1005 potentially explorable concept pairs we identified 120 unique concepts of 257 medically motivated concepts that were a subset of 623 unique concepts of the Medical Academic Word List.

TABLE I. THE HIGHEST-RANKING 3018 NOUNS IN BNC AND THE MOST FREQUENT WORD FAMILIES (I.E. CONCEPTS) OF MAWL AS WELL AS THE CONCEPTS OF MAWL AMONG 3018 NOUNS OF BNC HAVING THE HIGHEST RANKING AMONG THE NOUNS OF BNC AND AMONG THE CONCEPTS OF MAWL.

Highest-ranking 3018 nouns in BNC		Most frequent word families (i.e. concepts) of MAWL		Concepts of MAWL among 3018 nouns of BNC having the highest ranking among the nouns of BNC		Concepts of MAWL among 3018 nouns of BNC having the highest ranking among the concepts of MAWL	
concept	occurrences per million words	concept	occurrences	concept	ranking position in BNC / in MAWL	concept	ranking position in MAWL / in BNC
time	1833	cell	4421	area	21/23	cell	1/325
year	1639	data	2226	period	91/36	clinic	5/1701
people	1256	muscular	2049	community	98/328	factor	8/268
way	1108	significant	2039	issue	105/187	method	9/215
man	1003	clinic	1598	process	106/17	protein	10/1095
day	940	analyze	1447	research	112/25	tissue	11/1430
thing	776	respond	1427	section	137/78	dose	12/1659
child	710	factor	1237	team	146/451	gen.	13/1068
mr	673	method	1209	role	169/32	gene	13/1068
government	670	protein	1122	range	174/34	process	17/106

TABLE II. THE HIGHEST-RANKING INTERMEDIARY TRAVERSED CONCEPTS AND THE HIGHEST-RANKING TRAVERSED LINKS IN THE SHORTEST PATHS.

Highest-ranking intermediary traversed concepts in the shortest paths			Highest-ranking traversed links in the shortest paths (links are bidirectional and the concepts of the link are shown in an alphabetic order)			
concept	occurrences	belongs to the set of 257 concepts of MAWL	conceptA	conceptB	occurrences	concepts belonging to the set of 257 concepts of MAWL
management	333	no	management	team	105	b
analysis	208	no	analysis	research	77	b
development	197	no	development	team	76	b
health	155	no	cell	theory	67	a&b
major	115	yes	primary	role	65	a&b
research	107	yes	act	section	61	b
primary	107	yes	cancer	research	61	a&b
test	95	no	current	issue	59	b
code	85	yes	major	role	56	a&b
project	84	yes	model	role	56	b

IV. DISCUSSION

Motivated by our previous work (Lahti 2015c) we suggest that generating a conceptual co-occurrence network between a set of concepts belonging to a medical vocabulary can enable development of new computerized personalized methods for knowledge management of medical data for diagnosis and treatment as well as to enable creation of new adaptive personalized health support tools. We now report some results about our current data set of health-related information. Besides these results we consider that an essential part of our research contribution comes from our efforts of promoting the open data movement by publishing our full data set as an open access resource that can hopefully facilitate the modeling of medical knowledge and the development of personalized health support tools in the future research (our data set is available at Lahti (2016b)). With Yen's algorithm (Yen 1971) that computes top k shortest loopless paths we generated the shortest connecting paths between the concept pairs belonging to the set of 1005 unique potentially explorable concept pairs in the conceptual co-occurrence network. We gained altogether 9501 shortest paths for the set of 1005 unique concept pairs thus having for each concept pair on average 9.45 alternative shortest paths of the shortest length. The length of the shortest paths was on average 2.98 link steps and the median value was 3 link steps. There were 512 shortest paths of 4 link steps (for 5 unique concept pairs), 8285 shortest paths of 3 link steps (for 600 unique concept pairs), 678 shortest paths of 2 link steps (for 374 unique concept pairs) and 26 shortest paths of 1 link step (for 26 unique concept pairs). This means on average 102.4 shortest paths per each concept pair of a 4-link-step path, 13.8 shortest paths per each concept pair of a 3-link-step path, 1.8 shortest paths per each concept pair of a 2-link-step path and 1 shortest path per each concept pair of a 1-link-step path.

In the shortest paths between 1005 unique potentially explorable concept pairs in the conceptual co-occurrence network we identified 18784 intermediary traversed concepts, including 1540 unique concepts belonging to the set of 3018 highest-ranking unique nouns of BNC and 257 unique concepts belonging to the set of 257 medically motivated concepts of MAWL. In the shortest paths between 1005 unique potentially explorable concept pairs in the conceptual co-occurrence network we identified 28285 traversed links, including 8583 unique

links (when considering each link bidirectionally so that it is represented with such a link in which the concepts of the link are shown in an alphabetic order). Table 2 shows the highest-ranking intermediary traversed concepts and the highest-ranking traversed links in the shortest paths between 1005 unique potentially explorable concept pairs. Among 28285 traversed links the links not containing the concepts belonging to the set of 257 medically motivated concepts of MAWL ten highest-ranking links were: risk¤management (13), human¤development (12), analysis¤discourse (11), blood¤whole (11), human¤management (11), name¤address (11), blood¤whisky (10), health¤trainee (10), health¤statistics (9) and system¤management (9). In the shortest paths between 1005 unique potentially explorable concept pairs in the conceptual co-occurrence network we identified 18784 traversed two-step link paths, including 14875 unique two-step link paths (when considering each link bidirectionally so that it is represented with such a link in which the concepts of the link are shown in an alphabetic order). Table 3 shows the highest-ranking traversed two-step link paths in the shortest paths between 1005 unique potentially explorable concept pairs, among all paths and those paths in which all concepts belong to the set of of 257 concepts of MAWL.

TABLE III. THE HIGHEST-RANKING TRAVERSED TWO-STEP LINK PATHS IN THE SHORTEST PATHS BETWEEN 1005 UNIQUE POTENTIALLY EXPLORABLE CONCEPT PAIRS, AMONG ALL PATHS AND THOSE PATHS IN WHICH ALL CONCEPTS BELONG TO THE SET OF OF 257 CONCEPTS OF MAWL.

Highest-ranking traversed two-step link paths (bidirectionally) in the shortest paths between 1005 unique potentially explorable concept pairs, among all paths					Highest-ranking traversed two-step link paths (bidirectionally) in the shortest paths between 1005 unique potentially explorable concept pairs, among the paths in which all concepts belong to the set of of 257 concepts of MAWL			
conceptA	conceptB	conceptC	occurrences	concepts belonging to the set of 257 concepts of MAWL	conceptA	conceptB	conceptC	occurrences
human	understanding	role	13	c	function	primary	role	10
name	address	issue	11	c	phase	stance	issue	6
blood	whisky	issue	10	c	contrast	phase	project	5
function	primary	role	10	a&b&c	donor	major	role	5
health	primary	role	10	b&c	protein	complex	process	5
health	trainee	role	10	c	theory	prospect	issue	5
tumour	primary	role	10	b&c	cancer	alternative	method	4
management	compliance	section	8	b&c	cancer	cell	theory	4
mm	pp	issue	8	c	cancer	foundation	project	4
promoter	model	role	8	c	cancer	research	project	4

TABLE IV. HIGHEST-RANKING TRAVERSED CONCEPTS ALONG THE RANDOM EXPLORATION PATH OF 200 000 TRAVERSED LINK STEPS IN THE CONCEPTUAL CO-OCCURRENCE NETWORK CONTAINING 2994 UNIQUE NOUNS OF BNC AND 82814 UNIQUE LINKS AS WELL AS IN THE CO-OCCURRENCE NETWORK OF MEDICAL CONCEPTS CONTAINING 257 CONCEPTS OF MAWL AND 1116 UNIQUE LINKS.

Highest-ranking traversed concepts along the random exploration path of 200 000 traversed links steps in the conceptual co-occurrence network containing 2994 unique nouns of BNC and 82814 unique links				Highest-ranking traversed concepts along the random exploration path of 200 000 traversed links steps in the co-occurrence network of medical concepts containing 257 concepts of MAWL and 1116 unique links	
among all 2994 nouns of BNC		among the set of of 257 concepts of MAWL		among the set of of 257 concepts of MAWL	
concept	occurrences	concept	occurrences	concept	occurrences
management	403	sex	219	protein	3633
black	363	research	215	gene	2695
health	354	process	205	major	2661
red	309	energy	201	cancer	2530
american	296	design	195	file	2438
human	287	software	191	organism	2143
no	285	file	182	research	2132
music	277	technology	177	therapy	2129
public	271	cancer	166	cell	2121
air	266	final	158	transport	2050

We generated a random exploration path of 200000 traversed link steps in the conceptual co-occurrence network containing 2994 unique nouns of BNC and 82814 unique pairs of nouns (i.e. 82814 unique links). Table 4 shows the highest-ranking traversed concepts among all 2994 nouns of BNC along the random exploration path (all 2994 unique nouns of BNC became traversed) and among the set of of 257 concepts of MAWL along the random exploration path (all 257 unique concepts of MAWL became traversed). We identified that among 82814 unique pairs of nouns there were 1116 such unique pairs of nouns that both of the nouns belonged to the set of of 257 concepts of MAWL, and 243 of 257 concepts actually emerged in these concept pairs. Thus we could form a separate co-occurrence network of medical concepts containing 243 unique concepts of MAWL and 1116 unique pairs of nouns (i.e. 1116 unique links). We generated a random exploration path of 200000 traversed link steps in the co-occurrence network of medical concepts. Table 4 shows the highest-ranking traversed concepts among the set of of 243 concepts of MAWL along the random exploration path (241 of 243 unique concepts became traversed; it turned out that 2 of 243 unique concepts were interconnected with each other but separate from the other 241 concepts).

TABLE V. HIGHEST-RANKING TRAVERSED LINKS ALONG THE RANDOM EXPLORATION PATH OF 200000 TRAVERSED LINK STEPS IN THE CONCEPTUAL CO-OCCURRENCE NETWORK CONTAINING 2994 UNIQUE NOUNS OF BNC AND 82814 UNIQUE LINKS AS WELL AS IN THE CO-OCCURRENCE NETWORK OF MEDICAL CONCEPTS CONTAINING 257 CONCEPTS OF MAWL AND 1116 UNIQUE LINKS.

Highest-ranking traversed links along the random exploration path of 200000 traversed links steps in the conceptual co-occurrence network containing 2994 unique nouns of BNC and 82814 unique links				Highest-ranking traversed links along the random exploration path of 200000 traversed links steps in the conceptual co-occurrence network containing 257 concepts of MAWL and 1116 unique links	
among all 2994 nouns of BNC		among the set of of 257 concepts of MAWL		among the set of of 257 concepts of MAWL	
link	occurrences	link	occurrences	link	occurrences
romance¤nature	13	intake¤energy	8	purchase¤option	257
hire¤boat	13	sequence¤structure	8	option¤purchase	257
training¤doctrine	12	clinic¤foundation	7	purchase¤profile	253
craft¤stitch	11	node¤element	7	profile¤purchase	253
raid¤storage	11	acid¤calcium	6	adult¤alternative	223
creature¤fellow	10	aid¤grant	6	bacterium¤strain	221
operator¤radio	10	alternative¤method	6	surgery¤laser	220
smoke¤pipe	10	bacterium¤organism	6	terminal¤region	220
world¤series	10	bias¤ratio	6	fusion¤membrane	220
flexibility¤choice	10	cell¤cycle	6	region¤terminal	220

In the random exploration path of 200000 traversed link steps in the conceptual co-occurrence network containing 2994 unique nouns of BNC and 82814 unique links we identified 75385 unique links that became traversed at least once. Please note that we consider each link bidirectionally so that it is represented with such a link in which the concepts of the link are shown in an alphabetic order. Table 5 shows the highest-ranking traversed links in the conceptual co-occurrence network of 2994 unique nouns of BNC and 82814 unique links among 2994 nouns of BNC along the random exploration path and among the set of of 257 concepts of MAWL along the random exploration path. Table 5 shows the highest-ranking traversed links in the co-occurrence network of medical concepts containing 243 unique concepts of MAWL and 1116 unique links among the set of of 243 concepts of MAWL along the random exploration path. 1114 of 1116 unique links became traversed at least once; it turned out that 2 of 243 unique concepts were interconnected with each other but separate from the other 241 concepts.

REFERENCES

Abbott, A. (1992). From causes to events: Notes on narrative positivism. Sociological Methods & Research, 20(4), 428-55.

Abell, P. (2004). Narrative explanation: An alternative to variable-centered explanation? Annual Review of Sociology, 30, 287-310.

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/

Ashutosh, J., Andrews, D., Fiksdal, A., Kumbamu, A., McCormick, J., Misitano, A., Nelsen, L., Ryu, E., Sheth, A., Wu, S., & Pathak, J.(2014). Comparative Analysis of Online Health Queries Originating From Personal Computers and Smart Devices on a Consumer Health Information Portal. Journal of Medical Internet Research, 16(7). http://www.ncbi.nlm.nih.gov/pmc/articles/PMC4115262/

Atkinson, N., Saperstein, S., & Pleis, J. (2009). Using the internet for health-related activities: findings from a national probability sample. Journal of Medical Internet Research, 11(1): e4. doi: 10.2196/jmir.1035. http://www.jmir.org/2009/1/e4/

Brown, P., Halász, S., Goodall, C., Cochrane, D., Milano, P., & Allegra, J. (2010). The ngram chief complaint classifier: A novel method of automatically creating chief complaint classifiers based on international classification of diseases groupings. Journal of Biomedical Informatics, 43(2), 268–272.

Califf, R. (2014). The Patient-Centered Outcomes Research Network: A National Infrastructure for Comparative Effectiveness Research. North Carolina Medical Journal, 75(3).

Chuang, M., Lin, C., Wang, Y., & Cham, T. (2010). Development of pictographs depicting medication use instructions for low-literacy medical clinic ambulatory patients. Journal of Managed Care Pharmacy, 16(5), 337-345.

Cohen, R., & Havlin, S. (2003). Scale-free networks are ultrasmall. Physical Review Letters 90(5):058701.

Coulter, A., Parsons, S., & Askham, J. (eds.) (2008). Where are the patients in decision-making about their own care? World Health Organiztion, Policy brief. Health Systems and Policy Analysis. WHO Regional Office for Europe and European Observatory on Health Systems and Policies. http://www.who.int/management/general/decisionmaking/WhereArePatientsinDecisionMaking.pdf

Curé, O., Maurer, H., Shah, N., & Pendu, P. (2015). A formal concept analysis and semantic query expansion cooperation to refine health outcomes of interest. BMC Medical Informatics and Decision Making, 15(Suppl 1):S8.

Densen, P. (2011). Challenges and opportunities facing medical education. Transaction of the American Clinical and Climatoligal Association 122, 48–58.

Ferrer i Cancho, R., & Solé, R. (2001). The small world of human language. Proc. of the Royal Society of London, B., 268, 2261-2265.

FAU Erlangen-Nürnberg (2015). The Google Web 1T 5-Gram Database - SQLite Index & Web Interface - Collocations.

http://corpora.linguistik.uni-erlangen.de/demos/cgi-bin/Web1T5/Web1T5_colloc.perl

Feldman, M., & Pentland, B. (2003). Reconceptualizing organizational routines as a source of flexibility and change. Administrative Science Quarterly, 48, 94-118.

Giddens, A. (1984). The Constitution of Society. University of California Press, Berkeley, CA, USA.

Goode, A., Sukthankar, R., Mummert, L., Chen, M., Saltzman, J., Ross, D., Szymanski, S., Tarachandani, A., & Satyanarayanan, M. (2008). Distributed Online Anomaly Detection in High-Content Screening. Proc. IEEE International Symposium on Biomedical Imaging (ISBI).

Hayes. G., Lee, C., & Dourish, P. (2011). Organizational routines, innovation, and flexibility: the application of narrative networks to dynamic workflow. International Journal of Medical Informatics, 80(8): e161-77. doi: 10.1016/j.ijmedinf.2011.01.005. http://www.gillianhayes.com/wp-content/uploads/2012/08/J11_IJMI_NarrativeNetworks.pdf

Henriksson, A. (2013). Semantic spaces of clinical text. Leveraging Distributional Semantics for Natural Language Processing of Electronic Health Records. Licentiate Thesis, Department of Computer and Systems Sciences, Stockholm University. https://www.diva-portal.org/smash/get/diva2:653288/FULLTEXT01.pdf

Higgins, O., Sixsmith, J., Barry, M., Dmegan, C. (2011). A literature review on health information seeking behaviour on the web: a health consumer and health professional perspective. European Centre for Disease Control, Stockholm, Sweden. http://www.ecdc.europa.eu/en/publications/Publications/Literature%20review%20on%20health%20information-seeking%20behaviour%20on%20the%20web.pdf.

Hsu, Y., & Kao, H. (2013). CoIN: a network analysis for document triage. Database, The Journal of Biological Databases and Curation, Volume 2013, article id bat076.

Huang, Y., Lin, H., Hsu, Y., & Lin, J. (2012). Using n-gram analysis to cluster heartbeat signals. BMC Medical Informatics and Decision Making 2012, 12:64.

Kalenscher, T., & Tobler, P. (2008). Interdisciplinary Perspectives on Decision Making: Introduction. Cognitive, Affective, & Behavioral Neuroscience, 8(4), 345-347.

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

Karlgren, J., Bohman, M., Ekgren, A., Isheden, G., Kullmann, E., & Nilsson, D. (2014). Semantic Topology. Proc. Conference on Information and Knowledge Management (CIKM 2014), Shanghai, China. http://gavagai.se/wp-content/uploads/2014/11/km0697-karlgren.pdf

Kummervold, P., Chronaki, C., Lausen, B., Prokosch, H., Rasmussen, J., Santana, S., Staniszewski, A., & Wangberg, S. (2008). eHealth trends in Europe 2005-2007: a population-based survey. Journal of Medical Internet Research, 10(4): e42. doi: 10.2196/jmir.1023. http://www.jmir.org/2008/4/e42/

Lahti, Lauri (2015c). Educational exploration along the shortest paths in conceptual networks based on co-occurrence, language ability levels and frequency ranking. Proc. E-Learn - World Conference on E-Learning, 19-22 October 2015, Kona, Hawaii, USA. Association for the Advancement of Computing in Education (AACE), Chesapeake, VA, USA. http://www.editlib.org/p/151985/ (Open access: http://urn.fi/URN:NBN:fi:aalto-201509294488)

Lauri, Lahti (2016b). Supplement to Lauri Lahti's conference article ”Evaluation of semantic dependencies in a conceptual co-occurrence network of a medical vocabulary”, to appear online at http://aaltodoc.aalto.fi.

Larkey, L., & Croft, W. (1996). Combining classifiers in text categorization. Proc. 19th annual international ACM SIGIR conference on Research and development in information retrieval, 289–297. ACM.

Latour, B. (2005). Reassembling the Social: An Introduction to Actor-Network Theory. Oxford University Press, Oxford, UK.

Leech, G., Rayson, P., & Wilson, A. (2001). Word frequencies in written and spoken English: based on the British National Corpus. Longman, London, United Kingdom. ISBN 0582-32007-0. (A companion web site: Frequency lists. Chapter 5: Rank frequency lists of words within word classes (parts of speech) in the whole corpus. List 5.1: Frequency list of nouns (by lemma). Online available

at: http://ucrel.lancs.ac.uk/bncfreq/flists.html and http://ucrel.lancs.ac.uk/bncfreq/lists/5_1_all_rank_noun.txt)

Marling, C., Montani, S., Bichindaritz, I., & Funk, P. (2014). Synergistic case-based reasoning in medical domains. Expert Systems with Applications 41, 249–259.

Mikolov, T., Le, Q., & Sutskever, I. (2013). Exploiting similarities among languages for machine translation. Proc. International Conference on Language Research (ICLR 2013), Scottsdale, Arizona, USA. http://arxiv.org/pdf/1309.4168v1.pdf

Newman, M. (2000). Models of the small world. Journal of Statistical Physics, 101(3/4), 819-841.

O'Malley, A., & Marsden, P. (2008). The analysis of social networks. Health Services and Outcomes Research Methodology, 8(4), 222–269.

Pajevic, S., & Plenz, D. (2009). Efficient network reconstruction from dynamical cascades identifies small-world topology of neuronal avalanches. PLoS Computational Biology 5(1): e1000271.

Payne, P. (2012). Chapter 1: Biomedical Knowledge Integration. PLoS Computational Biology, 8(12), e1002826.

Pentland, B., & Feldman, M. (2008). Designing routines: On the folly of designing artifacts, while hoping for patterns of action. Information and Organization, 18(4), 235-50.

Pentland, B., & Feldman, M. (2008). Designing routines: On the folly of designing artifacts, while hoping for patterns of action. Information and Organization, 18(4), 235-50.

Qazvinian, V., & Radev, D. (2012). A Computational Analysis of Collective Discourse. Proc. Collective Intelligence conference 2012, 18-20 April 2012, Massachusetts Institute of Technology.

Ruttenberg, A., Clark, T., Bug, W., Samwald, M., Bodenreider, O., Chen, H., Doherty, D., Forsberg, K., Gao, Y., Kashyap, V., Kinoshita, J., Luciano, J., Marshall, M., Ogbuji, C., Rees, J., Stephens, S., Wong, G., Wu, E., Zaccagnini, D., Hongsermeier, T., Neumann, E., Herman, I., & Cheung, K. (2007). Advancing translational research with the Semantic Web. BMC Bioinformatics, Volume 8, Suppl 3.

Sadasivam, R., Kinney, R., Lemon, S., Shimada, S., Allison, J., & Houston, T. (2013). Internet health information seeking is a team sport: analysis of the Pew Internet Survey. International Journal of Medical Informatics, 82(3), 193–200. doi: 10.1016/j.ijmedinf.2012.09.008.

Sarkar, K. (2013). A Hybrid Approach to Extract Keyphrases from Medical Documents. International Journal of Computer Applications, 63(18).

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

Stanfill, M., Williams, M., Fenton, S., Jenders, R., & Hersh, W. (2010). A systematic literature review of automated clinical coding and classification systems. Journal of the American Medical Informatics Association, 17(6), 646–651. U.S. National Library of Medicine (2016). Medical Subject Headings (MeSH). https://www.nlm.nih.gov/mesh/meshhome.html

Stratton, P., & Wiles, J. (2010). Self-sustained non-periodic activity in networks of spiking neurons: The contribution of local and long-range connections and dynamic synapses. NeuroImage 52, 1070-1079.

Sulakhe, D., Balasubramanian, S., Xie, B., Berrocal, E., Feng, B., Taylor, A., Chitturi, B., Dave, U., Agam, G., Xu, J., Börnigen, D., Dubchak, I., Gilliam, T., & Maltsev, N. (2014). High-throughput translational medicine: challenges and solutions. Advances in Experimental Medicine and Biology, 799, 39-67.

Tripathy, B., Acharjya, D., & Cynthya, V. (2011). A framework for intelligent medical diagnosis using rough set with formal concept analysis. International Journal of Artificial Intelligence & Applications 2(2).

Wang, J., Liang, S., & Ge, G. (2008). Establishment of a Medical Academic Word List. English for Specific Purposes, 27, 442–458. http://ecourse.uoi.gr/pluginfile.php/93039/mod_resource/content/0/medical_academic_word_list.pdf

Wang, T., Wongsuphasawat, K., Plaisant, C., & Shneiderman, B. (2011). Extracting insights from electronic health records: Case studies, a visual analytics process model, and design recommendations. Journal of Medical Systems, 35(5), 1135–1152.

Wangberg, S., Andreassen, H., Prokosch, H., Santana, S., Sorensen, T., & Chronaki, C (2008). Relations between Internet use, socio-economic status (SES), social support and subjective health. Health Promotion International, 23(1), 70–7. doi: 10.1093/heapro/dam039published.

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

Workman, T., Rosemblat, G., Fiszman, M., & Rindflesch, T. (2013). A Literature-Based Assessment of Concept Pairs as a Measure of Semantic Relatedness. Proc. American Medical Informatics Association (AMIA) 2013, 1512–1521.

Yanardag, P., & Vishwanathan, S. (2015). Deep Graph Kernels. Proc. 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 1365-1374. http://web.ics.purdue.edu/~ypinar/kdd/deep_graph_kernels.pdf

Yen, J. (1971). Finding the k shortest loopless paths in a network. Management Science, 17(11), 712-716.

Zhou, X., Menche, J., Barabasi, A., & Sharma, A. (2014). Human symptoms-disease network. Nature Communications, 5, 4212. http://www.nature.com/ncomms/2014/140626/ncomms5212/abs/ncomms5212.html