Neurological Disorders and Publication Abstracts Follow Elements of Social Network Patterns When Indexed Using Ontology Tree-Based Key Term Search

Anand KULANTHAIVEL[a,b,1], Robert P. LIGHT[a], Katy BÖRNER[a], Chin Hua KONG[a], Josette F. JONES[b]

akulanth@indiana.edu[1], lightr@indiana.edu, katy@indiana.edu, kongch@indiana.edu, jofjones@iupui.edu

[a]*Indiana University Bloomington (Cyberinfrastructure for Network Science, Information & Library Science), Bloomington 47405 USA*

[b]*Indiana University-Purdue University Indianapolis (BioHealth Informatics), Indianapolis 46202 USA*

Abstract. Disorders of the Central Nervous System (CNS) are worldwide causes of morbidity and mortality. In order to further investigate the nature of the CNS research, we generate from an initial reference a controlled vocabulary of CNS disorder-related terms and ontological tree structure for this vocabulary, and then apply the vocabulary in an analysis of the past ten years of abstracts (N = 10,488) from a major neuroscience journal. Using literal search methodology with our terminology tree, we find over 5,200 relationships between abstracts and clinical diagnostic topics. After generating a network graph of these document-topic relationships, we find that this network graph contains characteristics of document-author and other human social networks, including evidence of scale-free and power law-like node distributions. However, we also found qualitative evidence for *Z*-normal-type (albeit logarithmically skewed) distributions within disorder popularity. Lastly, we discuss potential consumer-centered as well as clinic-centered uses for our ontology and search methodology.

Keywords: Ontology, information retrieval, neuroscience, networks, indexing, knowledge gaps, semantic medicine, translational medicine, knowledge discovery, neurology, psychiatry

1 Introduction

Research in the field of biomedical science associating publications with explicit clinical diagnostic terms is lacking. While central nervous system (CNS) disorders are a major cause of morbidity and mortality worldwide, there have been no studies to date on correlates between clinical and basic neuroscience terminology.

[1] Corresponding Author.

Given a controlled vocabulary (CV) whose members are organized into a tree-structured ontology, it is possible to search for biomedical or clinical meaning in a corpus of abstracts or other publication identifiers [1, 2]. If such an analysis is performed, one result may be the return of another ontology (this time, document-to-topic). The properties of such a network, as with any network, may be explored using basic social graph metrics [3].

Degree-based centrality (connectedness) is one measure of the influence of a node. Distributions of node centralities (including node degrees) have been postulated to allow conclusions to be drawn about a network in general given its centrality distributions [4]; Barabasi [5] in particular states that social-like network distributions, such as the power law distribution, are seen in a variety of situations that extend beyond sociology. Milojevic [6] proposes that modifications of power laws are allowed, including modification resulting in a Pareto Type-2 distribution. Finally, ranks among degree centralities of entities are proposed by Frasco et al [7] to have a geographic-social basis, possibly one laid on the foundations of individual-level interactions.

Therefore, a computational study of neuroscience publications with respect to clinical topics is likely to yield useful clues as to the *aboutness* of these publications and also reveal any topic bias(es). In particular, *Brain Research*, one particularly influential neuroscience journal with over 55,000 publications to date [8], provides an exemplar for a publication network of CNS-related topics. In this study, the past ten years (2004-2013) of abstracts written for *Brain Research* are analyzed against clinical terms from the Merck Manual for Professionals (in particular, the sections on neurological [9] and psychiatric [10] disorders).

2 Materials & Methods

2.1 Ontology Construction

The ontology and controlled vocabulary used in this study was derived from the Merck Manual for Professionals, particularly the sections on CNS pathologies [9, 10]. From this source, we found ninety-six (96) unique disorders. Each disorder super-heading was made into one diagnostic entity. Using the discretion of the authors, discrete and exclusive key words and key terms were mapped to each diagnostic entity. Therefore, a structure of disorder-to-keywords was created for each disorder. One example of a disorder-keywords tree as used here is seen in Figure 1.

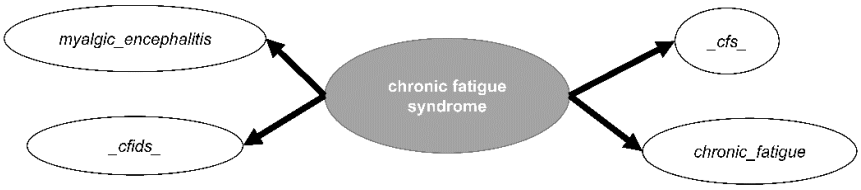

Fig. 1. Ontology map visualized (example; one of 96). Chronic fatigue syndrome is the diagnosis, and the entities it points to are the machine-searchable key terms that this diagnosis maps to.

2.2 Querying & Basic Information Retrieval

Information retrieval was performed by using the query term Brain Research[journal] in PubMed [11]. In order to represent the most recent cohort of documents, the search was filtered to only include articles published from 2004-2013. The result set was downloaded in XML format, and a raw corpus data file created using the Python scripting language. The output generated by the Python script formatted the corpus as PMID,abstract where PMID is the PubMed Identifier (PMID) for each article and abstract represents the abstract text of the article. Furthermore, in order to enhance machine parsing, all punctuation within abstract texts was removed and replaced with the underscore (_) symbol.

A separate file was created in order to contain the ontology trees. The format for each individual disorder tree was *disorder:key_term_1,key_term_2*, with each disorder having one or more key terms. For example, the ontology tree visualized in the above Figure 1 would have been represented in our search word file as *chronic_fatigue_syndrome: _cfs_,_cfids_,chronic_fatigue,myalgic_encephalitis*. Note the underscores surrounding acronyms; these are used to exclude words that might contain these strings as substrings. Underscores were also utilized to facilitate disambiguation of actual words that were less than four characters in length.

2.3 Parsing & Knowledge Synthesis

In order to create a graph-like representation of our subject-object construction (and in turn, discover which abstracts were related to which disorders), the PMID/abstract output file was searched against the ontology tree file, and positive matches sent to output in a network tool-readable edge list.

For this purpose, we wrote a custom program in Java (Virtual Machine; JVM) using the Eclipse IDE software tool [12]. The search algorithm utilized was literal, searching explicitly through the corpus file for disorder key words. As output, the algorithm generated an edge list file, with each line being an edge, the left node being the PMID, and the right node being the disorder topic that the matching key term was mapped to in the ontology tree file.

Of remark is that our algorithm was able to avoid parallel edges while constructing an edge list: Should the above document have contained cfs, cfids, and chronic fatigue, our algorithm will only output the pair 99999999,chronic_fatigue_syndrome once. This referential integrity was enforced by creating a step where JVM would store the previous keyword term match and refuse to generate a duplicate edge if the previous key term's parent diagnosis matched any other key term while the algorithm was searching for key terms of that particular diagnosis in that particular abstract.

2.4 Graph Preparation

The graph was prepared for diagramming as an undirected, unweighted network. In addition, the graph, while not explicitly bipartite as stated in the edge list, was of a bipartite topology (refer to Figures 2 and 3 for details).

2.5 Graph Visualization & Metrics

The Sci2 software tool (v1.1b) [13] was utilized for initial visualization and metrics computing. Specifically, the DrL layout within Sci2 [14] was used in order to gravitate the node positions for better viewing. Sci2 was then used to compute degree centrality measures for publications and disorders.

Correlations between degree measures were performed by exporting Sci2-generated data tables into Microsoft Excel [15] and analyzing and plotting the data in Excel for the histograms as well as for the disorder degree-rank scatter plot. Power analysis and regressions were performed in SAS v9.4 [16].

3 Results & Conclusions

3.1 Match Rates

Match Rate of Publications. Recall from our abstract that we searched 10,488 papers (i.e., the result set returned from Brain Research as [journal] term in PubMed, with the range set to past 10 years, and papers that only have available abstracts). 5,269 relationships were established between topics and publications. We noted that 4,163 papers (39.7%) had abstracts that matched the key terms in our ontology, yielding a corresponding miss rate of 60.3%. However, due to the limited terminology set of the current ontology tree, we chose to use graph analysis for the sub-corpus of publications whose abstracts did match our tree.

One must nonetheless realize that the low match rate, despite our best efforts in engineering the ontology for matching documents, may point to a disconnect (or knowledge gap) between science and medicine. On the other hand, there exists the possibility that many studies are carried out in order to study the normal functioning in the CNS (as opposed to disorder or pathology).

Disorder-Terminology Match Rate. Out of the 96 disorders found in the CNS section of the Merck Manual, 68 of these matched with publications via our ontology tree, yielding a disorder match rate of 70.8%.

3.2 Network Visualization

Network visualization yielded 9 graph components, with the giant component holding 99.6% of all nodes. Therefore, only the giant component is visualized in Figure 2.

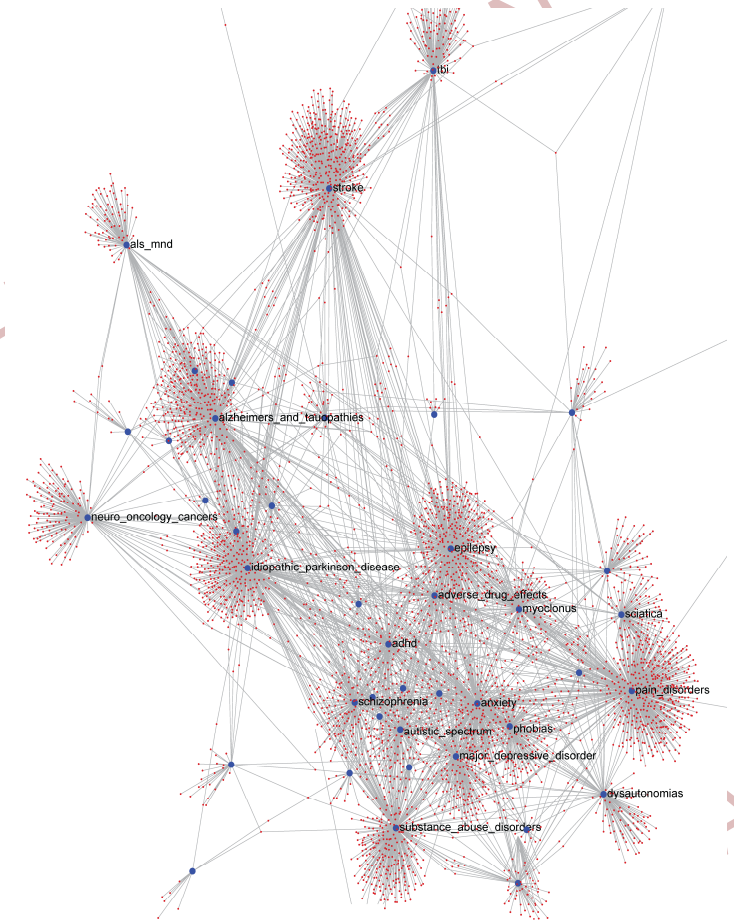

Fig. 2. Graph visualization of a majority of network giant component, performed using the DrL algorithm in the Sci2 Tool. Some high-profile disorders (Degree Centrality > 100) are highlighted by visible text labels. Most remaining disorders are highlighted by slightly larger node circles.

Remarks on Visual Topology. In our network graph (Figure 2), it is visually clear that a relatively small proportion of disorders studied held a wide amount of publication attention, while most disorders held relatively little attention. Some disorders in this layout appear to cluster via having many shared publications. For example, it is clearly

visible in Figure 2 that the entity stroke is well linked with tbi (traumatic brain injury). This particular linkage is viewed more closely in Figure 3.

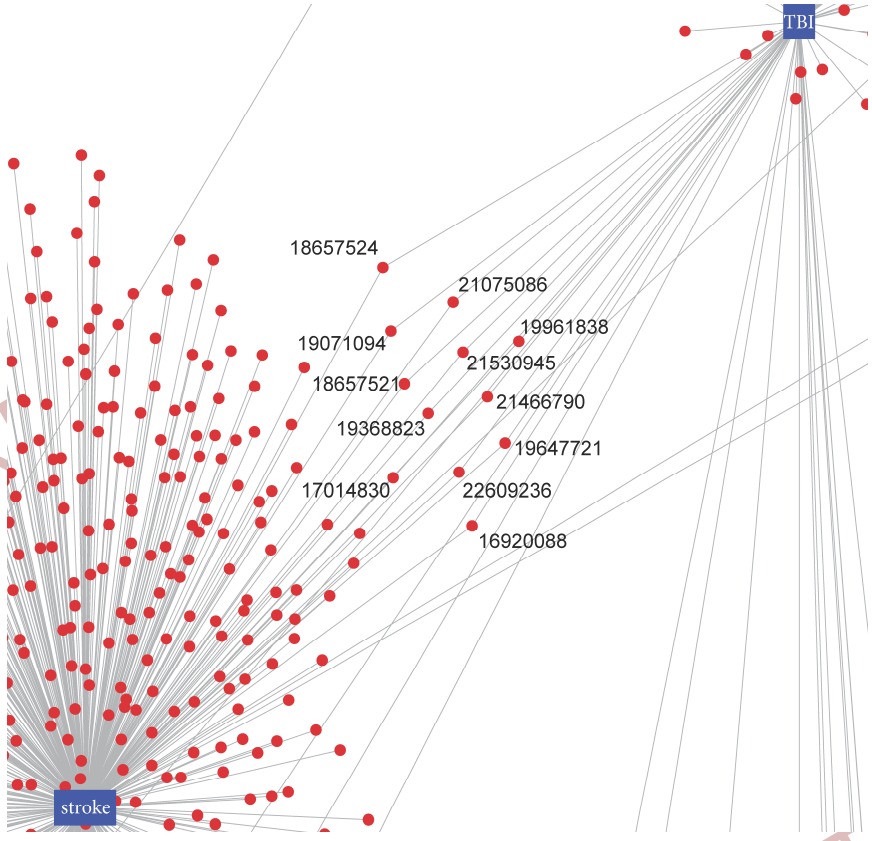

Fig. 3. Close-up view of two disorders our visualization and algorithm imply to be closely related. In this example, we see the entities *stroke* and *tbi* (traumatic brain injury) linked to each other by multiple publications, noted by their PMIDs.

Nonetheless, the linkage model must be viewed with some degree of suspicion, as stroke was found to be linked to 482 publications and TBI to 143. Therefore, while it was possible in theory to have 143 mutual matches between the two disorders, only 12 were observed, as seen in Figure 3.

3.3 Results of Graph Analysis

Regression of Ranking: Disorder Degree Centrality. In order to confirm the hypothesis of logarithmic distribution upon disorder degree rank, we transformed degree by log (base 10) function and then regressed against degree rank.

SAS returned a single variable power of disorder degree of > 0.999 at a = 0.05 for the series of disorder degrees. Single-factor ANOVA resulted in an F-value of 6710.15 (df = 66), and Pr > F was < 0.0001. The R-squared (RSQ) value for the regression was 0.990. P values for the slope and intercept of regression were both < 0.0001 (t stat = 132.19 for intercept and t stat = -81.92 for slope, respectively). Such a strong fit in the context of this linear-logarithmic transformation suggests a Pareto Type-2 distribution [6] and elements of Barabasi's theory of scale-free deterministic distributions [5] within the knowledge domain (neuroscience) at study.

Visualization and Plotting of Disorder Degree Rankings. The degree (number of publications connected to) of each of the discovered disorders we chose is plotted against the degree rank of these disorders in Figure 4; the Pareto-Type II distribution is more easily seen in the plot in Figure 4.

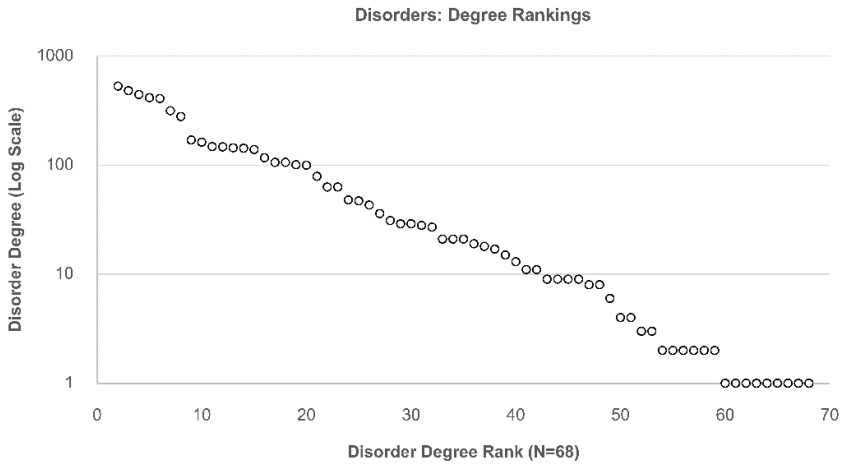

Fig. 3. Disorder degree-to-rank correlations. Rank is plotted on the X-axis, while disorder degree (connectedness to publications) is plotted on the Y-axis. The Y-axis is scaled logarithmically.

Qualitative Analysis of Disorder Degree Frequency Shows Elements of Long-Tailed and Natural Log-*Z*-Normal Distributions. Furthermore, when the publication frequency of disorders was binned for histogram analysis, the most parsimonious fit to a typical Z-type distribution appearance was obtained not by using linearly-sized bins, but instead, bins sized to powers of e (Euler's number; i.e., exponential binning was used). However, in our e-power histogram, there is still a peak of papers in bins corresponding to zero and one matches. This relationship is visualized in Figure 5. Topics of unusually high degree included pain disorders (D = 530 publications), stroke (D = 482 publications), and anxiety disorders (D = 279 publications).

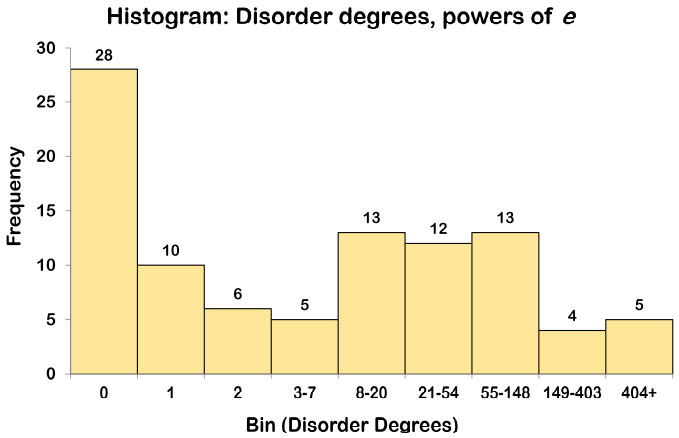

Fig. 4. Exponentially-binned histogram (powers of Euler's number) of disorder-to-publication degree distribution, shown on a linear frequency axis. Powers are rounded to the nearest whole number(s). A bin for unmatched disorders (degree = 0) is also included for reference on the far left of this figure.

Qualitative Analysis of Publication Degree Frequency Shows a Logarithmic-Linear Distribution. Similarly, degree for each publication was analyzed (i.e., how many disorders each publication would connect to). It appears that most publications were connected to only one disorder, while a few yielded matches with several disorders. The highest number of disorders matched for any abstract was six, with three abstracts matching six disorders each. A logarithmically-scaled histogram shows the linear-logarithmic trend in decreasing frequency of publications with respect to increasing topic degrees (Figure 6).

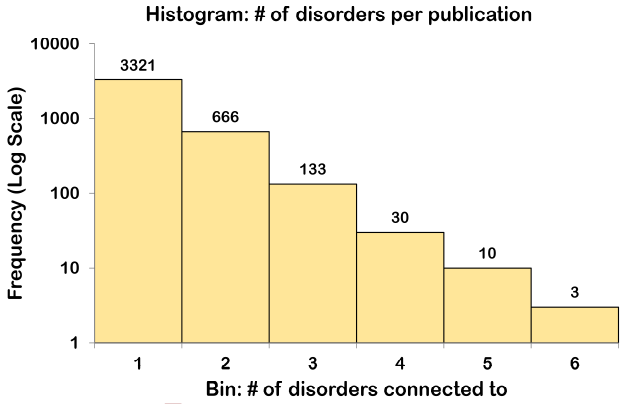

Fig. 5. Histogram of publication-to-disorder degrees (linear bin scale, logarithmic frequency scale). If non-matches were included in this histogram as they were included in Figure 5, another bin of magnitude 6,315 would be present to the left of the first bin.

4 Analysis & Discussion

4.1 Potential Reasons for Low Match Rate

Of great concern is that our model failed to match 60.3% of Brain Research abstracts. At this stage, we can only hypothesize upon the reasons for the lack of matching; a relatively small ontology (with only 260 base terms representing 96 disorders) could be of fault; the ontology must, in our opinions, be widened. The idea of full-text searching is also not to be excluded. Furthermore, there may exist large number of publications that do not explicitly describe CNS disorders per se, but normal functioning of the CNS; these publications would therefore evade any classification of disorders. Knowledge may also be gained by searching for related laboratory-to-clinic terminology; one example that was already used in the authors' ontology tree was the association of the terms nociceptor and nociception with pain disorders. Further discussion of improvements to the ontology is discussed in Section 5 of this report.

It also cannot be ignored that 28 out of 96 (i.e., 29.2%) of disorders, as mapped to their key terms with our ontology, did not match any of the publications. Along with the strong evidence for scale-free distributions by ranking [7], these non-matched disorders should be analyzed for epidemiological rates to determine if there exists true author-based or otherwise sociological bias against such disorders.

4.2 Distorted Distributions & Social Phenomenon

We required non-linear regression in topic ranking studies and exponential binning for frequencies in our publication topic distribution models; thus, our data supports some aspects of supporting social network patterning [5-7].

Classic co-authorship infometrics studies (ones that link papers to authors in similar networks) have shown that there exists` preferential attachment, that is, authors will preferentially attach to other authors who have had prior success publishing [6, 17]. Such co-authorship networks usually show power law (or power law-type) distributions and rankings of node degree, including the Pareto Type II distribution seen in our disorder-to-publication network. It follows, that we can properly speculate there is a preferential attachment of publications (and their authors) to certain topics. The ranking model (Figure 4) showed strong evidence for this hypothesis; we may certainly postulate from our data that neuroscience researchers tend to attach to established disorders, and quite possibly, to each other given the human-social foundations of attachment proposed by Frasco et al [7]. Nonetheless, we see that there was a great degree of specialization and again the potential for a scale-free [5] distribution. Such conclusion is supported in Figure 6, the drop in topics covered by any single publication is a steep logarithmic curve.

Ramifications of a Partial Log-*Z*-Normal Distribution in Disorder Degree. However, the explicit frequency model of disorder degree analysis (Figure 5) showed elements of both a log-*Z*-normal distribution (with peaks coinciding with values of 8 to

148 publications per disorder forming a bell curve-type shape in the exponential distribution in Figure 5) and a long-tailed [6] distribution, with the initial peak consisting zero and one matched publications. The former bell-like peak shows that there is a concentration of disorder study in more modest disorders, particularly given that the frequencies are scaled in a linear-normal fashion despite exponential binning.

5 Future Directions

5.1 Planned Studies: This Network; Improvements to Controlled Vocabulary

It is very important to note that the ontology (specifically, that of disorder-key term(s)) has not been curated by medical practitioners who deal with the CNS. We wish to subject the aforementioned ontology to validation by a panel of expert clinicians and researchers, possibly with a classical expert index card sort [18]. Such validation is likely to result in modest but significant modifications to this network based on explicit search terminology used.

We then wish to evaluate the ability of the revised network to draw conclusions on the interactions of humans with clinical information using a human survey project that will record the opinions of clinicians and researchers as they pertain to their beliefs on the importance of their own sub-fields of neuroscience and neurology. Finally, we intend on allowing consumers of healthcare (i.e., the lay public) to interact with this network map and discover how it changes (or reinforces) their perceptions of particular CNS disorders.

5.2 Study of Non-pathological CNS Function

A high non-match rate between our controlled vocabulary and the corpus of abstracts warrants further searching; we may in the future create a node entity of non-pathological, assign it key terms as we did with the 96 disorders, and re-perform our network visualization and analyses.

5.3 Recommendations: Use of Ontology and Algorithm as a Framework

With various ontologies commonly used as a framework in various information science applications, it is clear that this ontology (or a revised version thereof) ought to be used as a framework for the future study of CNS disorders. While we have only applied our ontology to relatively recent articles from *Brain Research*, studies of the resulting network over time (e.g., by comparison to similar networks generated for other publication time periods) would be of great interest. Furthermore, the ontology may be applied outside of *Brain Research* for the purpose of engineering knowledge from any corpus of documents that are CNS-related.

Potential for a Disorder Similarity Network. By viewing the links of publications between two given disorders, one may speculate as to how closely they are related (please refer back to Figure 4). For such similarity scores to be valid, however, we would require more disorder terminology (i.e., a higher N) for better publication match rates.

Creating Frameworks for Consumer Studies and Consumer Applications. As implied in Section 5.1, this ontology may help create a framework specifically for health consumer studies, engineering consumer-centered knowledge of the basic research sciences. It is also possible that such ontology may be useful in the context of electronic medical records (EMRs) for semantic analysis of consumers' self-reported health information in order to extract information regarding potential disorders that may be of concern to the consumers and their clinicians.

References

1. Skusa, A., Ruegg, A., Koehler, J.: Extraction of Biological Interaction Networks From Scientific Literature. Briefings in Bioinformatics. 6(3), 264-276 (2005)
2. Spasic I., Ananiadous S., McNaught J., Kumar A. Text Mining and Ontologies in Biomedicine: Making Sense of Raw Text. Briefings in Bioinformatics. 6(3), 239-251 (2005)
3. Yan, E., Ding, Y., Milojevic, S., Sugimoto, C.R.: Topics in Dynamic Research Communities: An Exploratory Story for the Field of Information Retrieval. Journal of Informetrics. 6(1), 140-153 (2012)
4. Newman, M.E.J.: Power Laws, Pareto Distributions, and Zipf's Law. Contemporary Physics 46(5), 323-351 (2005)
5. Barabasi, A.L., Erzsebet, R., Vicsek, T.: Deterministic Scale-Free Networks. Physica Acta. 299, 559-564 (2001)
6. Milojevic, S.: Power-Law Distributions in Information Science – Making the Case for Logarithmic Binning. Journal of the American Society for Information Science and Technology. 61(12), 2417-2425 (2010)
7. Frasco, G.F., Sun, J., Rozenfeld, H.D., ben-Avraham, D. Spatially-Distributed Social Complex Networks. Physical Review X. 4(011008) (2014).
8. Brain Research. (Journal). Information retrieved from http://journals.elsevier.com/brain-research/ on 19 December, 2013.
9. Merck & Co., Inc.: Neurological Disorders. (Section). In The Merck Manual of Diagnosis and Therapy for Professionals. (2013-2014). Retrieved from http://www.merckmanuals.com/professional/neurologic_disorders.html on 09 August 2013.
10. Merck & Co., Inc.: Psychiatric Disorders. (Section). In The Merck Manual of Diagnosis and Therapy for Professionals. (2013-2014) Retrieved from http://www.merckmanuals.com/professional/psychiatric_disorders.html on 09 August 2013.
11. United States Government, National Institutes of Health, National Library of Medicine (n.d.-2014): PubMed (Database Website). Retrieved from http://www.pubmed.gov/ (n.d.)

12. Eclipse Foundation, Inc.: Eclipse IDE for Java EE Developers (Software). Retrieved from http://www.eclipse.org/downloads/
13. Sci2 Team: Sci2 Tool (Software). (2009) Retrieved from http://sci2.cns.iu.edu/
14. Martin, S., Brown, W.M., Klavans, R., Boyack, K.W.: DrL: Distributed Recursive (Graph) Layout. SAND Reports. 2936, 1-10 (2008)
15. Excel 2014 (Software). Microsoft Corporation (Redmond/Seattle, Washington). Retrieved from http://www.microsoft.com/
16. SAS 9.4 (Software). SAS Institute (Cary, North Carolina).
17. Milojevic, S., Sugimoto, C.R., Yan, E., Ding, Y.: The Cognitive Structure of Library and Information Science: Analysis of Article Title Words. Journal of the American Society for Information Science and Technology. 62(10), 1933-1953 (2011)
18. Smith-Jentsch, K.A., Cannon-Bowers, J.A., Tannenbaum, S.I., Salas, E.: Guided Team Self-Correction: Impacts on Team Mental Models, Processes, and Effectiveness. Small Group Research. 39(3), 303-327 (2008)