

Hypothesis Generation for Joint Attention analysis on Autism

Jian Xu¹, Ying Ding², Chaomei Chen³, Erjia Yan³
¹ issxj@mail.sysu.edu.cn

School of Information Management, Sun Yat-sen University, Guangzhou, Guangdong (China)

² dingying@indiana.edu

Department of Information and Library Science, Indiana University, Bloomington, Indiana

³ chaomei.chen@drexel.edu and erjia.yan@drexel.edu

College of Computing and Informatics, Drexel University, Philadelphia, Pennsylvania

Introduction

Every 20 minutes a new case of autism is diagnosed worldwide, which affects around 6% of the population of children. One of the major challenges in autism is how to reliably diagnose autism as early as possible so that early intervention can be imposed to dramatically change the whole situation, even lead to cure. Joint attention is among these early impairments that distinguish young kids with autism from normal kids. Joint attention is a transdisciplinary area which was studied in robotics, psychology, autism, and neuroscience. However, Due to the unaware of similar or related researches in different domains, researchers are unknowingly duplicating studies that have already been done elsewhere. On the other hand, due to the lack of domain knowledge in other domains, researchers can experience difficulties to understand the advances in other domains. To deal with this dilemma, generating hypotheses is considered a potentially effective way. It is a crucial initial step for scientific breakthroughs, and usually relies on prior knowledge, experience and deep thinking. Especially for transdisciplinary domains, generating hypothesis from literature in different but related disciplines can be exciting and highly demanded because it is no longer possible for domain experts in one domain to fully master the knowledge in another domain.

Although marked with several decades of research history, it is until recent years that hypotheses generating attracts more attention in transdisciplinary research domains. Swanson (1986) proposed ABC model to inference the literature-based hypotheses. Later on, Srinivasan (2004) presented open and closed text mining algorithms that are built within the discovery framework established by Swanson and Smallheiser (reference here). Their algorithms successfully generated ranked term lists where key terms representing novel relationships between topics are ranked high. Zhang et al. (2014) established the semantic Medline which biomedical entities and association are semantically annotated using concepts in UMLS. They assumed that the network motifs in the network can represent basic interrelationships among diseases, drugs and genes and reflect a framework in which novel associations can be derived as hypotheses to be further validated by domain experts. Spangler et al. (2014) presented a prototype system KnIT, which can mine the

information contained in the scientific literature and represent it explicitly in a queriable network, and then further reason upon these data to generate novel and experimentally testable hypotheses. They applied their method to mine the publications related to p53 (a protein tumor suppressor) and are able to identify new protein kinases that phosphorylate p53. Malhotra et al. (2013) proposed a pattern matching approach for the detection of speculative statements in scientific text that uses a dictionary of speculative patterns to classify sentences as hypothetical. Their application on the domain of Alzheimer's disease showed that the automated approach captured a wide spectrum of scientific speculations and derived hypothetical knowledge leads to generation of a coherent overview on emerging knowledge niches. Song et al. (2007) constructed a Gene-Citation-Gene (GCG) network of gene pairs implicitly connected through citation and indicated that the GCG network can be useful for detecting gene interaction in an implicit manner.

In this initiative, we use text mining approach to analyze related publications on joint attention from robotics, psychology, autism and neuroscience, to generate hypotheses which will be tested in the lab which collects eye contact and body movement sensor data. Here some preliminary results were reported and discussed.

Methodology

Due to the transdisciplinary character of "joint attention" research, we elaborately selected eight data sources (Wiley Online Library, ProQuest PsycINFO, Science Direct, Scopus, Web of Science, PubMed Central, Springer Link and Google Scholar) to maximize the coverage of the final dataset. Phrase "joint attention" is used to search separately on each data source. Under the different download limitations, there are totally 39,845 records downloaded and 6,660 records left after remove duplicate records by the field "title". In the next step, keywords of each article in the dataset were extracted by using TF-IDF method. Then based on Keywords and other fields such as "journal name" and "citations", clustering were processed and relations among different clustering were analysed. By drawing the overall "research topic map", we can easily distinguish hot topics and their connections, and get to know their locations

on the overall map. Then different dimensions (e.g., age, speech, language, and communication) were defined to analyse the distribution of current researches. Finally, from different dimension analysis aspects, research blind points were uncovered and new hypotheses were inferred, which will be tested in the lab.

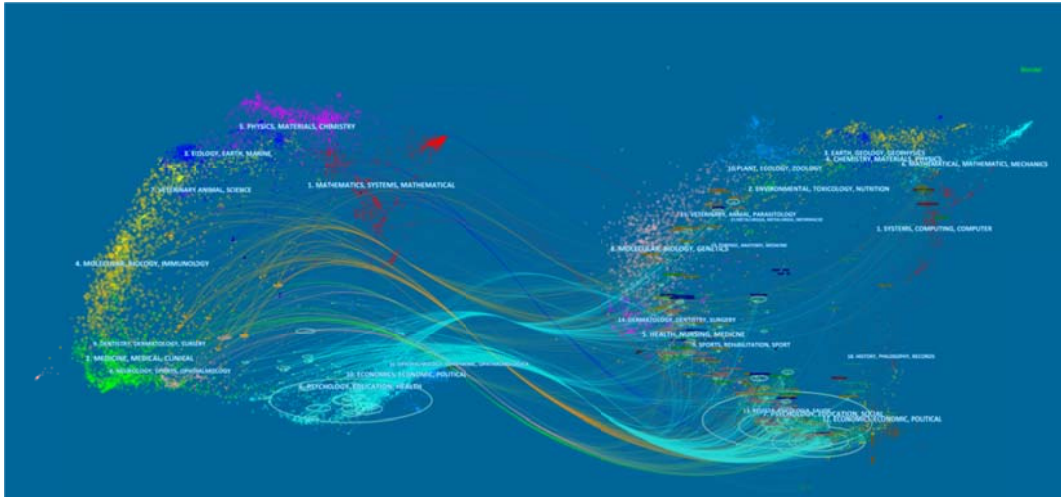


Figure 1: A dual-map overlay of "joint attention" search result from Web of Sciences.

Figure 1 shows the distribution of citing papers (left part) and cited papers (right part). Visualizations at this level are between journals, journal clusters, and overall maps. From the citation distribution and clustering results, we can identify the overall distribution of relevant sources and the most relevant targets (both ends with reference arcs). The label clustering result shows that the most popular domain discussing "joint attention" are Psychology, Education, Health, Medicine, Molecular, Economics, Mathematics, and Biology. It suggests that the Web of Science data is overwhelmingly dominated by a single journal *Journal of autism and developmental disorders*, with 169 papers. On the cited side, it is also the most cited journal in the dataset (6,640 citations). Other highly cited journals include *Child Development* (3,581 cites) and *Developmental Psychology* (2,328 cites).

Conclusions

This paper reports the ongoing effort on generating hypotheses in the transdisciplinary area of the joint attention research. We downloaded data from 8 separate data sources to maximize the coverage of "joint attention" related researches. Then text mining and visualization approaches were used to analyze related publications. Later stages of this research will generate hypotheses which will be tested in the lab based on current research distributions on different predefined dimensions.

Preliminary results

We tested a Web of Science query of "joint attention" (1,479 records) as a single dual-map overlay (Figure 1).

References

- Swanson, DR. (1986). Fish oil, Raynaud's syndrome, and undiscovered public knowledge. *Perspect Biol Med*, 30(1), 7–18.
- Srinivasan, P. (2004). Text mining: Generating hypotheses from MEDLINE. *Journal of the American Society for Information Science and Technology*, 55(5), 396-413.
- Zhang, Y., Tao, C., Jiang, G., Nair, A.A., Su, J., et al. (2014). Network-based analysis reveals distinct association patterns in a semantic Medline-based drug-disease-gene network. *Journal of Biomedical Semantics*, 5:33.
- Spangler, S., Wilkins, A.D., Bachman, B.J., Nagarajan, M., Dayaram, T., et al. (2014). Automated hypothesis generation based on mining scientific literature. *20th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD '14)* (pp. 1877-1886). New York:ACM
- Malhotra, A., Younesi, E., Gurulingappa, H., Hofmann-Apitius, M. (2013) 'HypothesisFinder: A Strategy for the detection of speculative statements in scientific text. *PLoS Comput Biol*, 9(7): e1003117.
- Song, M., Han, N.G., Kim, Y.H., Ding, Y., Chambers, T. (2014) Correction: Discovering Implicit Entity Relation with the Gene-Citation-Gene Network. *PLoS ONE*, 9 (1).
- Chen, C., Leydesdorff, L. (2014) Patterns of connections and movements in dual-map overlays: A new method of publication portfolio analysis. *Journal of the American Society for Information Science and Technology*, 65(2), 334-351.