Unlocking Knowledge Inheritance of Behavioral Research through Ontology Learning: An Ontology-Based Search Engine

Jingjing Li
jl9rf@comm.virginia.edu
University of Virginia

Kai R. Larsen
Kai.larsen@colorado.edu
University of Colorado at Boulder

Ahmed Abbasi
abbasi@comm.virginia.edu
University of Virginia

Demo/Prototype

The accumulated literature base in the behavioral sciences represents the most significant source of knowledge about human behavior, yet the same literature has grown beyond human comprehension, resulting in a knowledge inaccessibility problem. Existing IT artifacts such as full-text search engines have not been able to address this issue and in fact, may have intensified it by both rendering low-precision search results and escalating confirmation biases. Following the design science research paradigm, we propose a novel design framework and an instantiation—*TheoryOn*—to unlock behavioral knowledge embedded in large-scale behavioral articles. Based on an *ontology learning layer cake* framework and the state-of-the-art text analytics, we implemented an automated process of extraction and assembly of behavioral theories through *hypothesis*, *variable*, and *variable-relationship extraction*, and developed an ontology-based search engine—TheoryOn—that allows researchers to directly search for *constructs* and *synonymous constructs*, *construct relationships*, *antecedents* and *consequents*, and to easily integrate related *behavioral theories*. We conducted a randomized experiment comparing four information-retrieval tasks for behavioral literature review between *TheoryOn* and EBSCOhost (a full-text search engine) among 38 IS and Management researchers. On average, we found that *TheoryOn* users are significantly better at retrieving relevant constructs, construct relationships, and theories, suggesting significant benefit of proposed design artifact.

Keywords: Ontology Learning, Behavioral theories, Search Engines, Text Analytics, Randomized Experiment, Design Science Research

Introduction

Behavioral researchers continually search for and develop theories to improve disciplinary understanding of key phenomena. For example, IS has developed tens of thousands of theories representing important contributions to real-world IS phenomenon, some of them receiving tens of thousands more citations (Abbasi et al. 2016). Paradoxically, the rich academic literature on human behavior has become expansive to the point of incognizance over the past decades (e.g., Kraemer and Dutton 1991; Weber 2012). Studies have shown that researchers remain largely unaware of the majority of research, especially outside their own discipline (Larsen and Hovorka 2012), but also within narrow research areas (Colquitt and Zapata-Phelan 2007; Larsen and Bong 2016; Larsen 2002). This knowledge inaccessibility issue could result in literature fragmentation––reinventing construct relationships or hypotheses already introduced by others, or to proposing contradictory findings across different studies, prevent the building of cumulative traditions, leave the research community vulnerable to rapid change, and accrue tremendous monetary and social costs (Alexander et al. 1991; Bong 1996).

Beyond the apparent reasons (i.e., sheer numbers of publications and the lack of available time for researchers to read through them), we argue that the existing IT artifacts, such as full-text search engines are characteristically limited, and thus, are incapable of solving, and may in in fact, worsen, the knowledge accessibility problem. Full-text search engines like Google Scholar and EBSCOhost have similar characteristics. They manage information at the article-level, provide keyword search of the free text in abstracts or full-texts, and incorporate paper-level citation analysis and usage statistics for the ranking of results (Beel and Gipp 2009). These characteristics

result in severe false positives in returned results (Boeker et al. 2013). For instance, a search for the construct, *perceived usefulness*, intended to represent the perceived belief that a system can enhance job performance (Davis 1989), will result in 808,000 returned results (as of 7/1/2016). Rather than the actual construct, *perceived usefulness*, many of the returned articles contain the loosely used phrase, *perceived usefulness*, or constructs carrying the same name but representing different latent concepts, such as Nelson's (1991) *perceived usefulness* scale that measures the perceived importance of skill proficiency on job performance.

Following the design science paradigm (Gregor and Hevner 2013; Hevner et al. 2004; Simon 1996), this study proposes a design artifact—an ontology-based search engine, named TheoryOn—to alleviate the knowledge inaccessibility problem in the behavioral sciences and to address the weaknesses of existing IT artifacts. We adopt Weber's (2012) view that a behavioral theory "accounts for some subset of phenomena in the real world" and is a specialized type of Bunge's (1977; 1999) ontology.[1] Therefore, we use the ontology learning layer cake (Buitelaar et al. 2005)—a process of extracting relevant parts of ontologies (i.e., concepts, relations, and axioms) from texts by using a collection of techniques and resources—as a *kernel theory* to guide our design process of extracting behavioral theories from existing, large-scale behavioral publications. We narrow our focus to a manageable initial level by focusing on behavioral positivist research, and specifically those fitting the criteria of Gregor's (2006) theories for explanation and theories for explanation and prediction (natural science types of research). We

[1] There may be alternative notions about the mapping between behavioral theories and ontologies, but this is not the focus of the paper. By adopting Weber's view, many of the ontology-learning tasks and techniques can be nicely adapted to guide extracting behavioral theories from a large-scale behavioral publication.

illustrate the usefulness of the proposed behavioral ontology learning layer cake by developing an instance, in this case an ontology-based search engine named *TheoryOn* that extracts *hypotheses, constructs,* and *theoretical relationships* from hundreds of relevant behavioral studies published at *MIS Quarterly*, *Information Systems Research*, and the *Journal of Applied Psychology*—all top journals in their fields (Li et al. 2016a; 2016b; 2017; 2018a; 2018b; 2019). With the extracted theory "parts," *TheoryOn* allows researchers to directly search for *constructs*, *construct relationships*, and *theoretically related constructs* (e.g. antecedents or consequents), as well as easily integrate *related theories*.

Behavioral Knowledge Search Needs

To solve the behavioral knowledge inaccessibility issue, we need to understand how researchers make use of behavioral knowledge—the search needs. We discover these needs by conducting semi-structured interviews with five behavioral researchers who are junior- to senior-level faculty researchers in the IS and organizational behavioral fields. As a result, we identified four academic search needs agreed upon by most of interviewees:

1) Construct search. Full-text search engines, such as Google Scholar and EBSCOhost, operate at the article level and does not extract theory-relevant meta-data (construct and construct relationships) embedded in articles, which leads to a large number of false positives in behavioral knowledge searches. Therefore, the proposed search engine should eliminate articles that contain phrases similar to construct names but are used as general phrases or constructs with identical names but with different meanings, and it should only return articles that contain the relevant constructs, including synonymous constructs that have the same meaning but under different

names. All of the interviewed researchers raised concerns about construct naming issues and demanded this function.

2) Construct relationship search. Searching by construct pairs enables researchers to quickly identify whether or not a proposed construct relationship is already being studied. However, it is not possible for full-text search engines to specify a search for any paper positing a relationship between construct X and construct Y because they operate at the article level.

3) Theoretically related construct search. We define a construct, C', to be theoretically related to another construct, C, if they both appear in the same hypothesized theoretical model, (e.g., antecedents, consequents, or control variables for a focal construct). All interviewees indicated such a need to quickly find theoretically related constructs to help them find control variables and theorize construct relationships.

4) Theory visualization and integration. Behavioral theories could be visualized as network graphs with constructs as nodes and relationships as arrows. This visualization format allows researchers to quickly skim through as many articles as possible and identify a comprehensive list of theories for further examination quickly. Additionally, it could facilitate inter-theory relationship comparison and building theory evolution graph (Mueller 2015). Moreover, it could facilitate theory integration where theory graphs can be connected by clustering synonymous constructs. Such integration is useful for nomological network construction (Lee et al. 2003) and meta-analysis (Larsen and Bong 2016). Three interviewees indicated a need to build a nomological network graph that integrates related theories.

Design of the Artifact

Considering a behavioral theory, consisting of constructs and their relationships, as a special type of ontologies, this study represents the first effort of extracting behavioral theories from large-scale behavioral publications using ontology learning layer cake framework. We propose a novel behavioral ontology learning framework[2] based on ontology learning to automate the extraction of hypotheses, variables, and their relationships from academic articles for eventual inclusion in search engines (Figure 1). The ontology learning for behavioral theory could be broken into five tasks: *hypothesis extraction*, *variable extraction and grouping*, *theoretical relationship extraction*, *construct hierarchy building*, and *theoretical relationship discovery*. Each task generates an output corresponding to ontology learning's five outputs: *terms*, *concepts*, *non-taxonomic relations*, *taxonomic relations*, and *axioms*, respectively (Fu et al. 2008; 2010; 2012).

Following the proposed ontology learning framework, we automatically extracted hypotheses, variables, and variable relationships from articles from MIS Quarterly, Information Systems Research and Journal of Applied Psychology in the period of 1990-2007. We then implemented a search engine, namely TheoryOn, with four functionalities (Figure 2) according to the presented four search needs. In the subsequent sections, we briefly describe these four functionalities (detailed information provided in the video links).

[2] The framework is based on the following assumptions: 1) an article belonging to the Gregor's theory for explanation and theory for explanation and prediction usually presents its behavioral theory through hypotheses; 2) a hypothesis usually includes a statement describing the relationships between variables; 3) the majority of hypotheses in the articles are supported, representing viable construct relationships in the theoretical models. Under conditions of even partial support for such assumptions, the drastic increases in available evidence about hypothesized relationships and theories are likely to make up for missed relationships and theories in articles that do not fit the assumptions.

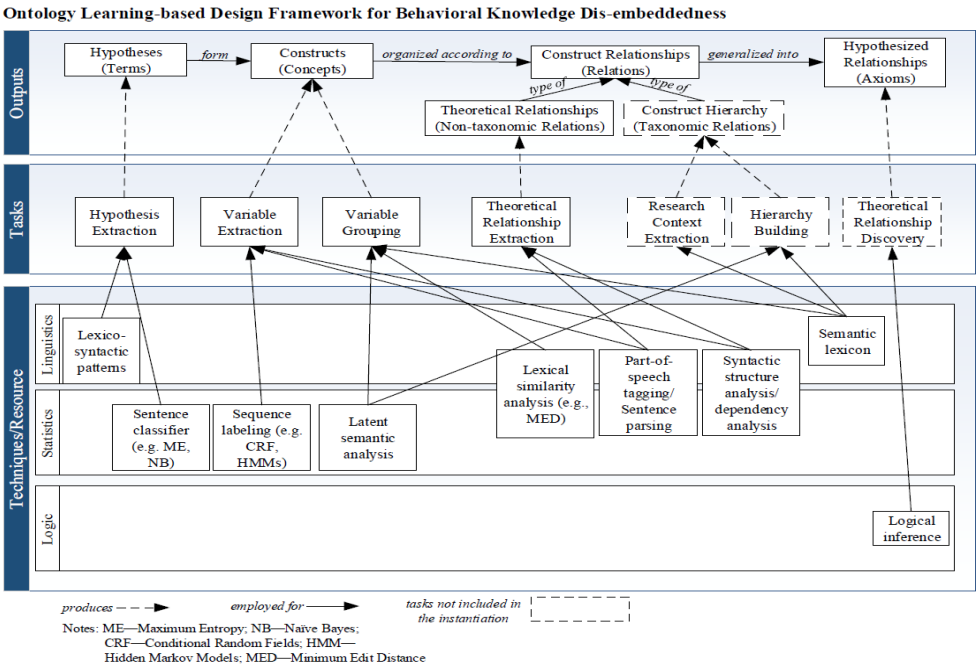
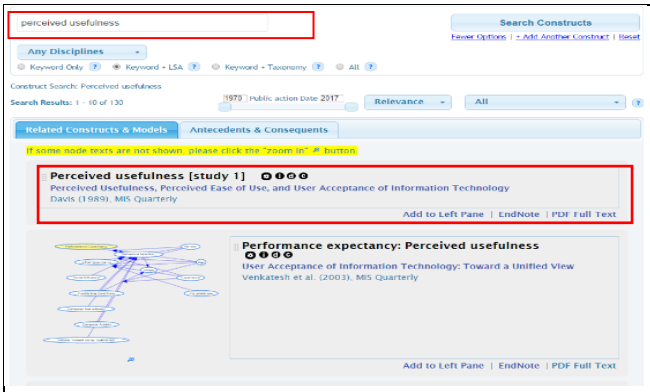
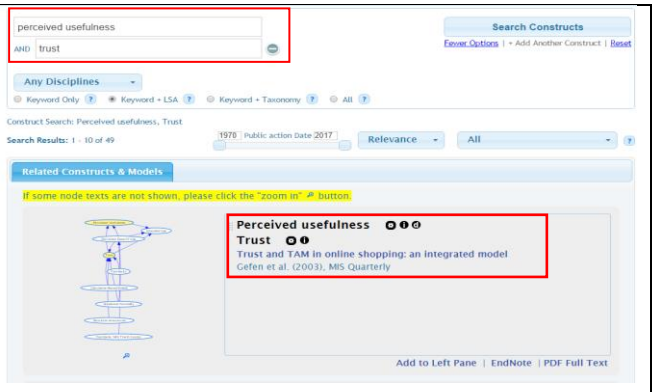


Figure 1. Proposed Design Framework for Behavioral Knowledge Dis-embeddedness.

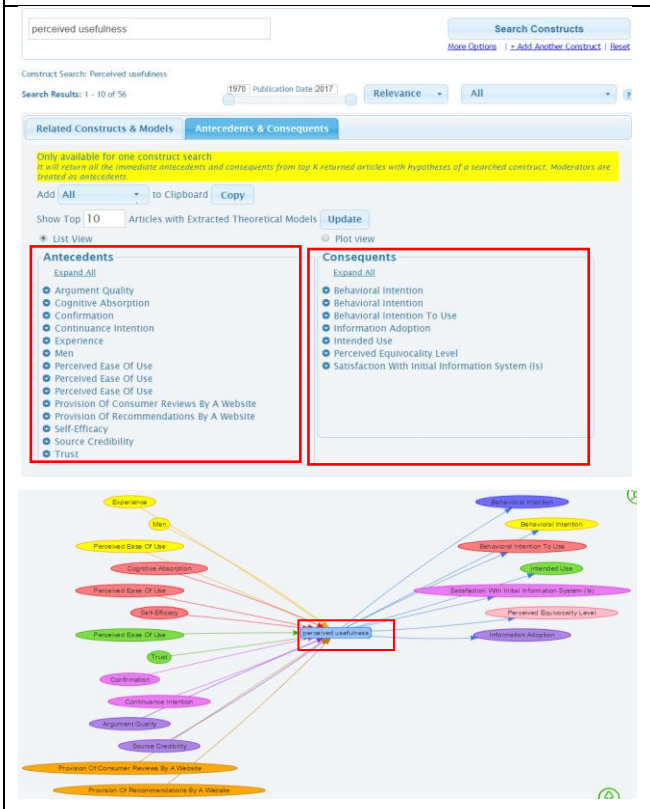
1) Construct Search. TheoryOn allows users to specify a construct in a search query and only return articles that containing this construct or its synonymous constructs. The construct information is directly presented in the returned results. Users can also save the related constructs and articles in a sorting hierarchy. For more details, watch the video "TheoryOn: Synonymous Construct Search".



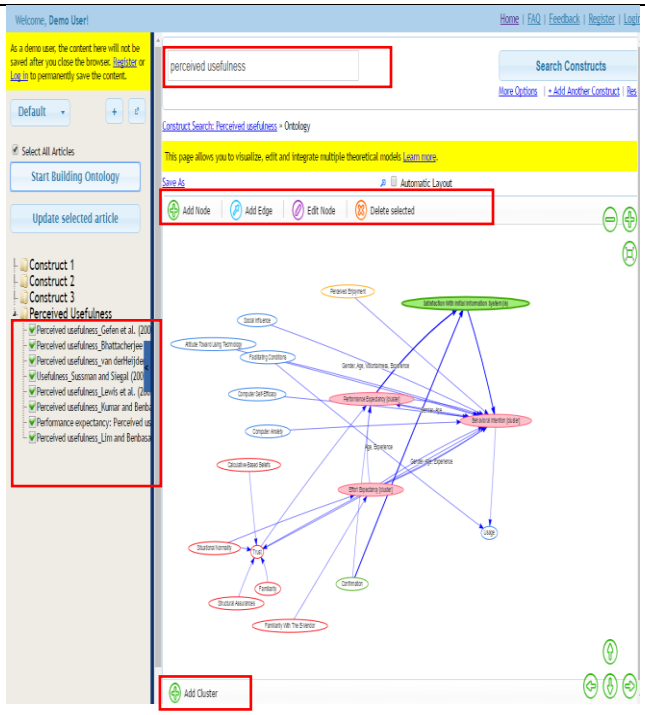
Construct Search



Construct-Pair Search



Theoretically Related Construct Search



Theory Integration

Figure 2. Screenshots of TheoryOn's Four Major Functionalities

2) Construct-Pair Search. TheoryOn allows users to specify a construct pair in a search query and only returns articles that containing these two constructs. The constructs and their

relationships are shown in the extracted theoretical models in the left part of the search results. For more details, watch the video "TheoryOn: Construct-Pair Search".

3) Theoretically Related Construct Search. A user could not only inspect the theoretical models containing a construct of interest (highlighted in yellow) but also examine its antecedents and consequents in a list or plot view. For more details, watch the video "TheoryOn: Theoretically Related Construct Search".

4) Theory Integration. All the related theories can be saved in the sorting hierarchy and visualized on the canvas. A user can then integrate theories by clustering synonymous constructs or customize the theoretical networks by editing, deleting or adding any nodes and links. For more details, watch the video "TheoryOn: Theory Integration".

Evaluation of the Artifact

To evaluate the usability and usefulness of TheoryOn, we conducted a randomized experiment with 38 Information Systems and Organizational Behavior Ph.D. students from a variety of programs in the U.S. and around the world. We designed four tasks that correspond to the four search needs—construct search, construct-pair search, antecedent and consequent search, and theory integration—and evaluated the performance of TheoryOn against the control group that used a common full-text search engine, EBSCOhost. Their behavioral information retrieval performances were compared using precision and recall (Salton 1989), which illustrate the tendency to reduce false positives and false negatives in an information retrieval task, respectively. On average, *TheoryOn* users were 16.84% better at precision and 72.66% better at recall. Additionally, we found that *TheoryOn* was perceived to be more useful and easier to use than

EBSCOhost, as evaluated by a Perceived Usefulness ($M_{theoryon}$ = 5.88, M_{ebsco} = 4.71, $p<0.01$) and an Ease of Use ($M_{theoryon}$ = 6.20, M_{ebsco} = 5.17, $p<0.01$) scales (7-point likert) adapted from *Unified Theory of Acceptance and Use of Technology* (UTAUT) proposed by (Venkatesh et al. 2003).

Significance of Research and Practice

Our contribution of this study is manifold. First, we propose an ontology learning design framework specific for behavioral research to guide incremental development of behavioral theory knowledge-management systems. Second, we instantiate the framework into an ontology-based search-engine artifact, namely *TheoryOn*, to show the applicability of the framework. Third, we outline a research agenda, or map, for the behavioral ontology learning research area. Finally, we demonstrate the value of a knowledge base for behavioral research findings through a randomized experiment. Overall, the knowledge contribution of this research represents an instance of *exaptation* in which we adapted solutions from the ontology-learning field to a new problem of extracting behavioral theories from large-scale behavioral publications , a novel combination of a problem and its solutions added to the Gregor and Hevner's examination of past literature in *MIS Quarterly* (2013). We believe the work has important implications for disembedding behavioral knowledge in various social science domains including IS and health (Netemeyer et al. 2019; Zahedi et al. 2015; Zimbra et al. 2010), including potential for predicting behavioral relationships (Brown et al. 2015a; 2015b). Our work also advances the state-of-the-art for natural language processing (Kitchens et al. 2018; Deng et al. 2018; Zimbra et al. 2018; Adjeroh et al. 2014; Benjamin et al. 2014) and text analytics (Abbasi et al. 2018a; 2018b; 2019; Ahmad et al. 2019; Khaja et al. 2018).

References

Abbasi, A., Sarker, S., and Chiang, R. H. L. "Big Data Research in Information Systems: Toward an Inclusive Research Agenda," *Journal of the Association for Information Systems*, 17(2), no. 3, 2016.

Abbasi, A., Li, J., Clifford, G. D., and Taylor, H. A, "Make 'Fairness By Design' Part of Machine Learning," *Harvard Business Review*, August 5, 2018a, digital article: https://hbr.org/2018/08/make-fairness-by-design-part-of-machine-learning

Abbasi, A. Zhou, Y., Deng, S., and Zhang, P. "Text Analytics to Support Sense-making in Social Media: A Language-Action Perspective," *MIS Quarterly*, 42(2), 2018b, pp. 427-464.

Abbasi, A., Li, J., Adjeroh, D., Abate, M., and Zheng W. "Don't Mention It? Analyzing User-generated Content Signals for Early Adverse Event Warnings," *Information Systems Research*, 30(3), 2019, pp. 1007-1028.

Adjeroh, D., Beal, R., Abbasi, A., Zheng, W., Abate, M., and Ross, A. "Signal Fusion for Social Media Analysis of Adverse Drug Events," *IEEE Intelligent Systems*, 29(2), 2014, pp. 74-80.

Ahmad, F., Abbasi, A., Li, J., Dobolyi, D., Netemeyer, R., Clifford, G., and Chen, H. "A Deep Learning Architecture for Psychometric Natural Language Processing," *ACM Transactions on Information Systems*, 2019, forthcoming.

Alexander, P. A., Schallert, D. L., and Hare, V. C. 1991. "Coming to Terms: How Researchers in Learning and Literacy Talk About Knowledge," *Review of educational research* (61:3), pp. 315-343.

Beel, J., and Gipp, B. 2009. "Google Scholar's Ranking Algorithm: An Introductory Overview," *Proceedings of the 12th International Conference on Scientometrics and Informetrics (ISSI'09)*: Rio de Janeiro (Brazil), pp. 230-241.

Benjamin, V., Chung, W., Abbasi, A., Chuang, J., Larson, C. A., and Chen, H. "Evaluating Text Visualization for Authorship Analysis," *Security Informatics*, 3(10), 2014.

Boeker, M., Vach, W., and Motschall, E. 2013. "Google Scholar as Replacement for Systematic Literature Searches: Good Relative Recall and Precision Are Not Enough," *BMC medical research methodology* (13:1), p. 131.

Bong, M. 1996. "Problems in Academic Motivation Research and Advantages and Disadvantages of Their Solutions," *Contemporary Educational Psychology* (21:2), pp. 149-165.

Brown, D. E., Abbasi, A., and Lau, R. Y. K., "Predictive Analytics: Predictive Modeling at the Micro Level," *IEEE Intelligent Systems*, 30(3), 2015a, pp. 6-8.

Brown, D. E., Abbasi, A., and Lau, R. Y. K., "Predictive Analytics," *IEEE Intelligent Systems*, 30(2), 2015b, pp. 6-8.

Buitelaar, P., Cimiano, P., and Magnini, B. 2005. *Ontology Learning from Text: Methods, Evaluation and Applications*. IOS press.

Bunge, M. 1977. *Treatise on Basic Philosophy Volume 3: Ontology I-the Furniture of the World*. Dordrecht, Holland: D. Reidel Publishing Company.

Bunge, M. 1999. *Treatise on Basic Philosophy: Volume 4: Ontology Ii: A World of Systems*. Dordrecht, Holland: D. Reidel Publishing Company.

Colquitt, J. A., and Zapata-Phelan, C. P. 2007. "Trends in Theory Building and Theory Testing: A Five-Decade Study of the Academy of Management Journal," *Academy of Management Journal* (50:6), pp. 1281-1303.

Davis, F. D. 1989. "Perceived Usefulness, Perceived Ease of Use, and User Acceptance of Information Technology," *MIS Quarterly* (13:3), pp. 319-340.

Deng, S., Zhang, P., Zhou, Y., and Abbasi, A. "Using Discussion Logic in Analyzing Online Group Discussions: A Text Mining Approach," *Information and Management*, 56(4), 2019, pp. 536-551.

Fu, T., Abbasi, A., and Chen, H. "A Hybrid Approach to Web Forum Interactional Coherence Analysis," *Journal of the American Society for Information Science and Technology*, 59(8), 2008, pp. 1195-1209.

Fu, T., Abbasi, A., and Chen, H. "A Focused Crawler for Dark Web Forums," *Journal of the American Society for Information Science and Technology*, 61(6), 2010, pp. 1213-1231.

Fu, T., Abbasi, A., Zeng, D., and Chen, H. "Sentimental Spidering: Leveraging Opinion Information in Focused Crawlers," *ACM Transactions on Information Systems*, 30(4), 2012, no. 24.

Gregor, S. 2006. "The Nature of Theory in Information Systems," *MIS Quarterly* (30:3), pp. 611-642.

Gregor, S., and Hevner, A. R. 2013. "Positioning and Presenting Design Science Research for Maximum Impact," *MIS Quarterly* (37:2), pp. 337-355.

Hevner, A., March, S., Park, J., and Ram, S. 2004. "Design Science in Information Systems Research," *MIS Quarterly* (28:1), pp. 75-105.

Khaja, H. I., Abate, M., Zheng, W., Abbasi, A., and Adjeroh, D. "Evaluating Semantic Similarity for Adverse Drug Event Narratives," *In the 11th International Conference on Social Computing, Behavioral-Cultural Modeling, and Prediction (SBP)*, Washington D.C., 2018.

Kitchens, B., Dobolyi, D., Li, J., and Abbasi, A. "Advanced Customer Analytics: Strategic Value through Integration of Relationship-Oriented Big Data," *Journal of Management Information Systems*, 35(2), 2018, pp. 540-574.

Kraemer, K. L., and Dutton, W. H. 1991. "Survey Research in the Study of Management Information Systems," in *The Information Systems Research Challenge: Survey Research Methods*, K.L. Kraemer (ed.). Boston, MA: Harvard Business School, pp. 3-57.

Larsen, K. R., and Bong, C. H. 2016. "A Tool for Addressing Construct Identity in Literature Reviews and Meta-Analyses," *MIS Quarterly* (40:3), pp. 529-551; A521-A521.

Larsen, K. R., and Hovorka, D. S. 2012. "Developing Interfield Nomological Nets," in: *Hawaii International Conference on System Sciences*. Maui, Hawaii: IEEE.

Larsen, K. R. T. 2002. "A Taxonomy of Antecedents to Is Success: Variable Analysis Studies," 2002-003, University of Colorado, Information Systems, Boulder, pp. 1-95.

Lee, Y., Kozar, K., and Larsen, K. R. T. 2003. "The Technology Acceptance Model: Past, Present, and Future," *Communications of the AIS* (12:50).

Li, J., Larsen, K., and Abbasi, A. "TheoryOn: Designing a Construct-based Search Engine to Reduce Information Overload for Behavioral Science Research," *In the 11th International Conference on Design Science Research in Information Systems and Technology (DESRIST)*, St. John's, Canada, May 23-24, 2016a.

Li, J., Abbasi, A., Cheema, A., and Abraham, L. "Path to Purpose? Impact of Online Purchases' Hedonic and Utilitarian Characteristics on the Customer Journey," *In the 26th Workshop on Information Technologies and Systems (WITS)*, Dublin, Ireland, December 15-16, 2016b.

Li, J., Larsen, K., and Abbasi, A. "Unlocking our Behavioral Knowledge Inheritance through Ontology Learning: A Design Framework, an Instantiation, and a Randomized Experiment," *In the INFORMS Workshop on Data Science*, Houston, TX, October 21, 2017.

Li, J., Abbasi, A., Ahmad, F., and Chen, H. "PyNDA: Deep Learning for Psychometric Natural Language Processing," *In the INFORMS Workshop on Data Science*, Phoenix, AZ, November 3, 2018a.

Li, J., Abbasi, A., Ahmad, F., and Chen, H. "Deep Learning for Psychometric NLP," *In the 28th Workshop on Information Technologies and Systems (WITS)*, Santa Clara, CA, December 15-16, 2018b.

Li, J., Larsen, K., and Abbasi, A. "TheoryOn: A Design Framework and System for Unlocking Behavioral Knowledge through Ontology Learning," *MIS Quarterly*, 2019, conditionally accepted.

Mueller, R. M. 2015. "A Meta-Model for Inferring Inter-Theory Relationships of Causal Theories," *System Sciences (HICSS), 2015 48th Hawaii International Conference on*: IEEE, pp. 4908-4917.

Nelson, R. R. 1991. "Educational Needs as Perceived by Is and End-User Personnel: A Survey of Knowledge and Skill Requirements," *MIS Quarterly* (15:4), pp. 503-525.

Netemeyer, R., Dobolyi, D., Abbasi, A., Clifford, G., and Taylor, H. "Health Literacy, Health Numeracy and Trust in Doctor: Effects on Key Consumer Health Outcomes," *Journal of Consumer Affairs*, 2019, forthcoming.

Salton, G. 1989. "Automatic Text Processing: The Transformation, Analysis, and Retrieval Of," *Reading: Addison-Wesley*).

Simon, H. 1996. *The Science of the Artificial (3rd Ed.)*. Cambridge, MA: MIT Press.

Taylor, H. A., Henderson, F., Abbasi, A., and Clifford, G. D., "Cardiovascular Disease in African Americans: Innovative Community Engagement for Research Recruitment and Impact," *American Journal of Kidney Diseases*, 72(5)(Supp 1), 2018, S43-S46.

Venkatesh, V., Morris, M. G., Davis, G. B., and Davis, F. D. 2003. "User Acceptance of Information Technology: Toward a Unified View," *MIS Quarterly* (27:3), pp. 425-478.

Weber, R. 2012. "Evaluating and Developing Theories in the Information Systems Discipline," *Journal of the Association for Information Systems* (13:1), pp. 1-30.

Zahedi, F. M., Abbasi, A., and Chen, Y. "Fake-Website Detection Tools: Identifying Design Elements that Promote Individuals' Use and Enhance their Performance," *Journal of the Association for Information Systems*, 16(6), 2015, pp. 448-484.

Zimbra, D., Abbasi, A., and Chen, H. "A Cyber-Archaeology Approach to Social Movement Research: Framework and Case Study," *Journal of Computer-Mediated Communication*, 16, 2010, pp. 48-70.

Zimbra, D., Abbasi, A., Zeng, D., and Chen, H. "The State-of-the-Art in Twitter Sentiment Analysis: A Review and Benchmark Evaluation," *ACM Transactions on Management Information Systems*, 9(2), 2018, no. 5.