

## **THEORYON: A DESIGN FRAMEWORK AND SYSTEM FOR UNLOCKING BEHAVIORAL KNOWLEDGE THROUGH ONTOLOGY LEARNING<sup>1</sup>**

**Jingjing Li**

McIntire School of Commerce, University of Virginia,  
Charlottesville, VA 22903 U.S.A. {jl9rf@comm.virginia.edu}

**Kai Larsen**

Leeds School of Business, University of Colorado,  
Boulder, CO 80309 U.S.A. {Kai.Larsen@colorado.edu}

**Ahmed Abbasi**

Mendoza College of Business, University of Notre Dame,  
Notre Dame, IN 46556 U.S.A. {aabbasi@nd.edu}

*The scholarly information-seeking process for behavioral research consists of three phases: searching, accessing, and processing of past research. Existing IT artifacts, such as Google Scholar, have in part addressed the searching and accessing phases, but fall short of facilitating the processing phase, creating a knowledge inaccessibility problem. We propose a behavioral ontology learning from text (BOLT) design framework that presents concrete prescriptions for developing systems capable of supporting researchers during their processing of behavioral knowledge. Based upon BOLT, we developed a search engine—TheoryOn—to allow researchers to directly search for constructs, construct relationships, antecedents, and consequents, and to easily integrate related theories. Our design framework and search engine were rigorously evaluated through a series of data mining experiments, a randomized user experiment, and an applicability check. The data mining experiment results lent credence to the design principles prescribed by BOLT. The randomized experiment compared TheoryOn with EBSCOhost and Google Scholar across four information retrieval tasks, illustrating TheoryOn's ability to reduce false positives and false negatives during the information-seeking process. Furthermore, an in-depth applicability check with IS scholars offered qualitative support for the efficacy of an ontology-based search and the usefulness of TheoryOn during the processing phase of existing research. The evaluation results collectively underscore the significance of our proposed design artifacts for addressing the knowledge inaccessibility problem for behavioral research literature.*

**Keywords:** Behavioral ontology learning design framework, design science research, text analytics, machine learning, randomized experiment, applicability check

### **Introduction**

Behavioral researchers continually search for and develop theories to improve disciplinary understanding of key phe-

nomena. For example, the theory of planned behaviors that explains an individual's intention to engage in a certain behavior has received more than 70,000 citations (Ajzen 1991). Hundreds of theories have been developed or extended (Soper and Turel 2015) to facilitate the understanding of real-world information systems phenomena, some receiving tens of thousands of citations (e.g., Davis 1989; Venkatesh et al. 2003). Paradoxically, though, the rich academic literature

<sup>1</sup>Jeffrey Parsons was the accepting senior editor for this paper. Ofir Arazy served as the associate editor.

on human behavior has become expansive to the point of incognizance over the past few decades (Weber 2012). Since behavioral research takes a concept-centric perspective, the completeness of any literature search is often defined as the proportion of relevant constructs retrieved (Webster and Watson 2002). In this regard, studies have shown that researchers remain largely unaware of the majority of research, especially outside of their own disciplines, but also within narrow research areas (Colquitt and Zapata-Phelan 2007). Larsen and Bong (2016) have shown that even for a small set of full-text articles, experts could retrieve, on average, fewer than 10% of the articles that would be valuable for a literature review and knowledge acquisition.

The result is knowledge inaccessibility in behavioral research, here defined as the situation that behavioral knowledge embedded in the extant large-scale literature may not be accessed by researchers in a comprehensive and accurate manner. Knowledge inaccessibility could have a considerable negative impact on behavioral research in at least four ways. First, with knowledge inaccessibility, researchers are prone to literature fragmentation and end up reinventing constructs, relationships, or hypotheses already introduced by others. This can result in wasted and redundant research efforts (Spell 2001), possible errors such as spurious gap-spotting and gap-patching, and the generation of marginal research (Rai 2017). Second, it prevents the building of cumulative traditions in which researchers build on each other's previous work and that "definitions, topics, and concepts are shared" (Keen 1980), thus threatening the development and progression of a research field (Im and Straub 2012). Third, it introduces inefficiencies in research processes and knowledge acquisition and construction, leaving the research community to be slow in accommodating emerging contexts (Quirchmayer et al. 2012), low in research topic agility (Peffers 2002), and vulnerable to rapid environmental change (Trinh et al. 2012). Finally, as behavioral research spans multiple disciplines, including medicine, psychology, sociology, education, and economics, impediments to the knowledge creation process and spurious research findings resulting from knowledge inaccessibility may exact tremendous monetary and social costs (Weber 2012).

The research on information-seeking behaviors in behavioral research could shed light on how knowledge inaccessibility arises. Unlike natural sciences whose theories consist of strictly universal statements and languages (Popper 1980), behavioral theories often measure beliefs, expectations, attitudes, and emotions through constructs and relationships defined by malleable and ever-changing language systems (Arnulf et al. 2014, 2018; Larsen et al. 2013). Hence, it is important to adopt a construct-centric view (Webster and Watson 2002) and clarify and synthesize construct relationships during the scholarly information search process.

Specifically, behavioral researchers' information seeking can be categorized into phases, including *searching*, *accessing*, and *processing* (Meho and Tibbo 2003), of which the *accessing* phase serves as a conduit between the critical *searching* and *processing* phases. Existing IT artifacts, such as full-text search engines, are well suited for the *searching* phase, in which the process of identifying relevant and potentially relevant materials is initiated. For instance, Google Scholar and EBSCOhost provide keyword searches of the free text in abstracts or full texts and incorporate article-level citation analysis and usage statistics for results ranking (Beel and Gipp 2010). However, the majority of knowledge inaccessibility issues manifest in the *processing* phase, which entails extraction, synthesis, and analysis of concepts across articles. High false-positive rates in full-text search engines due to lack of behavioral knowledge extraction can mislead researchers into prematurely ending the information-seeking process (Boeker et al. 2013). False negatives, demonstrated as confirmation biases (White 2013), could also hinder the completeness of the *processing* outcomes. Specifically, confirmation biases occur as a result of individual researchers' and research fields' proclivity toward "unwitting selectivity in the acquisition and use of evidence" (Nickerson 1998, p. 175). These biases are amplified by full-text search engines' keyword matching and data-dependent ranking algorithms. In this sense, complementing full-text search engines with new search artifacts capable of disembedding behavioral knowledge to better support the *processing* phase may help enhance information-seeking abilities.

To alleviate the knowledge inaccessibility problem, this paper proposes two design artifacts: a behavioral ontology learning from text (*BOLT*) design framework and an ontology-based search engine, *TheoryOn*, to disembed behavioral knowledge from existing, large-scale publications. Using relevant behavioral research (Baron and Kenny 1986; Larsen and Bong 2016; Larsen et al. 2019; Weber 2012), our *BOLT* design framework views behavioral knowledge as a specialized type of ontology whose core parts include *hypotheses*, *constructs*, and *construct relationships*. Hence, effective behavioral ontology learning entails appropriate extraction of these parts. Referring to ontology learning from text literature (e.g., Wong et al. 2012) and the pertinent natural language processing (NLP) research, *BOLT* identifies the underlying tasks and prescribes the best potential techniques. We further used the proposed design framework to develop the ontology-based search engine *TheoryOn*, which allows researchers to directly search for *constructs*, *construct relationships*, and *theoretically related constructs*, as well as to easily integrate *related theories*. We also conducted a multifaceted evaluation of *TheoryOn* (Gill and Hevner 2013; Hevner et al. 2004) which included ontology learning method and system comparison experiments, a randomized user experiment comparing it with

the EBSCOhost and Google Scholar search engines, and an applicability check. Overall, the contribution of our work represents an instance of *exaptation* in which we adapted solutions from the ontology learning field to a new problem: disembedding behavioral knowledge from large-scale behavioral publications (Gregor and Hevner 2013).

## Background: Limitations of Existing Search Engines to Support Scholarly Literature Review

At a high level, behavioral researchers' information seeking can be categorized into three closely inter-related phases: *searching*, *accessing*, and *processing* (Ellis 1989; Meho and Tibbo 2003). *Searching* encompasses "the period where identifying relevant and potentially relevant materials is initiated" (Meho and Tibbo 2003, p. 584). This phase includes steps such as initial search, following chains of citations, and casually browsing selected articles (Ellis 1989). *Processing* is where synthesizing and analyzing across articles and concepts takes place (Meho and Tibbo 2003). This phase is especially important for behavioral research because, unlike natural sciences, behavioral research measures beliefs, perceptions, and emotions that are less amenable to being described in universal languages. Hence, there is a greater need to scrutinize, differentiate, filter, organize, and amalgamate information across articles. Since information seeking in behavioral research is a nonlinear process, the *accessing* phase simply serves as a conduit between the critical *searching* and *processing* phases. These phases are consistent with information-seeking stages identified through our survey of IS scholars (see Appendix D for details).

Full-text academic search engines such as Google Scholar and EBSCOhost are especially well suited to supporting the *searching* phase. They allow individual users to specify search queries that represent their search needs and return as search results a subset of articles that contain all or some keywords from the search query (Beel and Gipp 2010). They also allow researchers to conduct keyword searches within articles that cited a relevant paper. The efficiency and ease of use of full-text search engines are ideal for the initial search phase, high-level browsing of potentially relevant articles, and quickly making sense of an area of research through query keyword expansion and citation network traversal. Conversely, full-text search engines are not as well suited to supporting the *processing* phase of information-seeking behavior, where false positive and false negative errors can adversely affect synthesizing and analyzing activities.

First, full-text search engines do not extract behavioral knowledge-relevant metadata embedded in articles (e.g., constructs

and construct relationships), which can lead to a large number of false positives in behavioral knowledge searches. For instance, a search for the construct *perceived usefulness*, intended to represent the perceived belief that a system can enhance job performance (Davis 1989), returned 90,200 results in Google Scholar (retrieved on January 31, 2019). Rather than the actual construct, *perceived usefulness*, most of the returned articles contained 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, which measures the perceived importance of skill proficiency on job performance. Boeker et al. (2013) found that across 14 existing systematic studies, the precision of Google Scholar was 0.13%. Similarly, Yousafzai et al. (2007) evaluated 36,463 articles in a Google Scholar search result for the technology acceptance model and found precision to be 0.39%, indicating that finding all relevant articles would require evaluating hundreds of false positives for each truly relevant article found.

Second, keyword matching and citation- and usage-based ranking, although efficient and effective in supporting the searching phase, may lead to heavy false negatives and amplify researcher- and field-level confirmation bias during the processing phase (Larsen et al. 2019). Search queries based on keywords confirm researchers' preexisting beliefs about construct names and research topics. Correspondingly, a search, for example, for the *subjective norm* construct may altogether miss articles that employ identical operationalizations, but with different names, such as *social factors* or *image*. This type of confirmation bias, originating from researchers' tendency to confirm existing beliefs while neglecting viable alternatives, is referred to as *researcher-level confirmation bias* (Bushman and Wells 2001). Furthermore, the citation- and usage-based results ordering (Beel and Gipp 2010), a typical function of search engines such as Google Scholar, amplify the researcher-level bias by promoting articles subject to confirmation bias to the top of the search results (Beel and Gipp 2010; White 2013), resulting in a field-level confirmation bias.

Admittedly, researchers and practitioners from scholarly information retrieval fields have proposed academic support IT artifacts that go beyond keyword-based indexing and allow direct search on metadata from academic articles. For example, Quan et al. (2004) applied a fuzzy formal concept analysis (FFCA) method to build a scholarly ontology from a citation database, an important step toward building an ontology-based search engine. Semanticscholar.org extracts authors, journals, conferences, figures, references, and topics from academic articles to facilitate a more nuanced search. Similarly, Microsoft Academic Search extracts seven entity types, including authors, affiliations, title, year, journal, conference series, and field of study to help users quickly process

knowledge embedded in academic articles. However, these systems focus on metadata such as authors, citations, and journals, and do not incorporate provisions for behavioral knowledge disembedding (e.g., hypotheses, constructs, and relationships), rendering them less effective for facilitating the processing phase of scholarly information-seeking and solving the behavioral knowledge inaccessibility problem.

Another related scholarly information retrieval field is biomedical text mining, which utilizes NLP techniques to extract genes, proteins, drugs, diseases, and their relations from the biomedical literature (e.g., Luo et al. 2016). However, biologists have the advantage of gene nomenclature committees (Eyre et al. 2006) and the good fortune of working with constructs—genes—that have more precise definitions and are more amenable to consistent measurement through commonly accepted instruments. Conversely, behavioral constructs are often defined by malleable and ever-changing language systems. Consequently, the methods suitable for biomedical text mining may be less so in behavioral knowledge disembedding.

## Design Framework for Disembedding Behavioral Knowledge

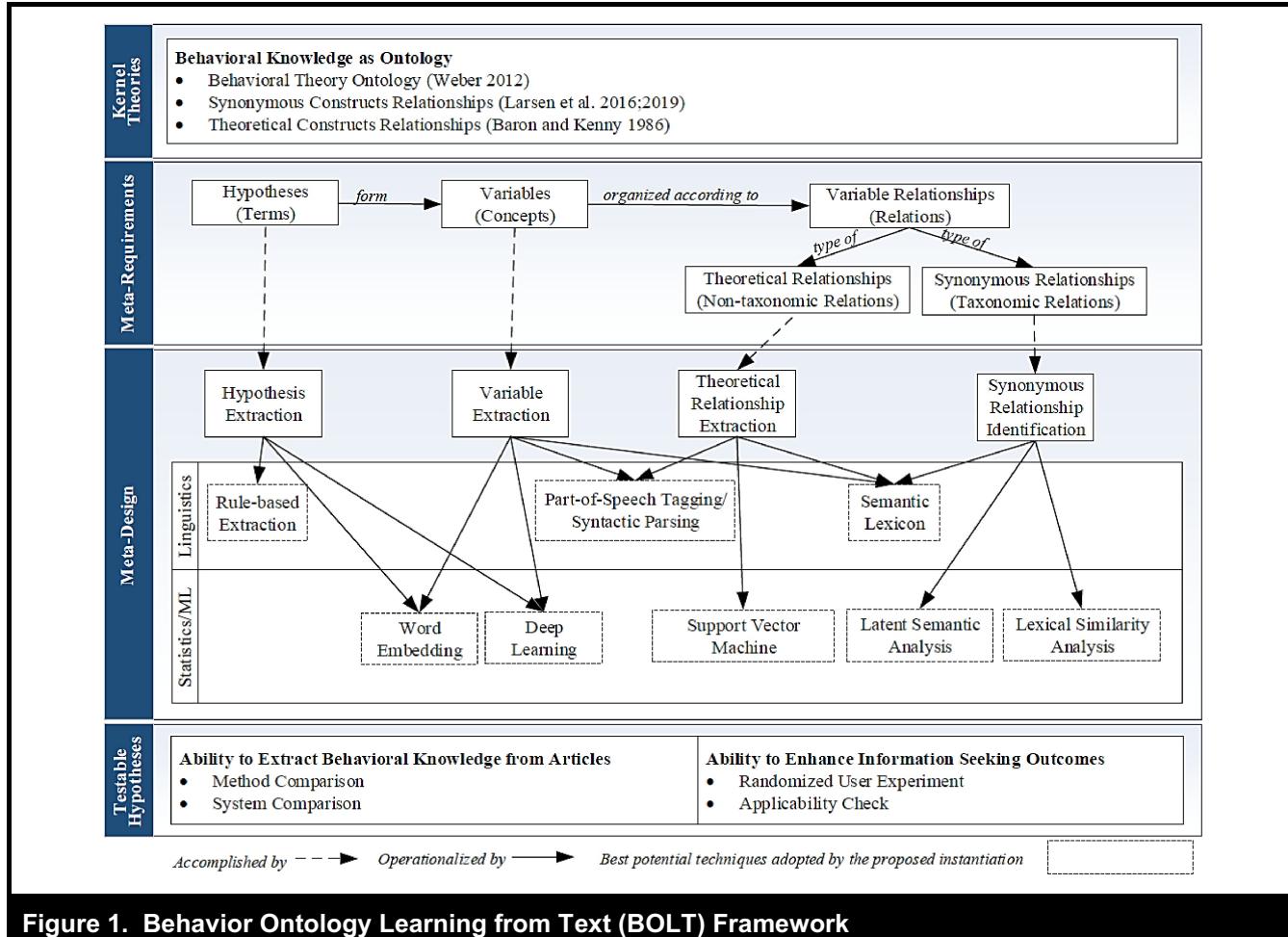
Given the need for disembedding behavioral knowledge to improve the scholarly information-seeking process and the lack of dedicated IT artifacts that address this need, an important set of questions arises. How do we define behavioral knowledge? What are the key features and capabilities that behavioral knowledge disembedding systems should support? What are the necessary tasks to accomplish them? Relative to keyword-based search engines, metadata systems, and biomedical text analytics tools, systems geared toward extracting behavioral knowledge face greater ambiguity. Behavioral research involves various philosophical paradigms, spans numerous disciplines, and employs a number of research methods, which lead to a considerable level of disagreement about what constitutes behavioral knowledge (e.g., Corley and Gioia 2011; Gregor 2006; Weber 2012). Consequently, there is lack of clarity on the key features and capabilities that should be supported by behavioral knowledge disembedding systems, not to mention the necessary tasks and techniques to accomplish them.

Design guidelines are needed due to this complexity of properly representing behavioral knowledge and the resulting ambiguity regarding the key features, capabilities, and tasks a behavioral knowledge disembedding system should support. In accordance with recent calls (e.g., Larsen et al. 2017; Larsen et al. 2020), we therefore propose a design framework

named behavioral ontology learning from text (BOLT), to guide the development of systems for *extracting behavioral knowledge encompassed in large-scale, multidisciplinary publication databases*. According to the design science paradigm (Hevner et al. 2004; Walls et al. 1992), design is both a product and a process. The design product concerns a set of requirements and design characteristics to guide IT artifact construction. Meanwhile, the design process involves steps and procedures to construct the IT artifact, and typically follows a highly iterative process consisting of building and evaluating (March and Smith 1995). Our design framework focuses on the design product, which is composed of *kernel theories*, *meta-requirements*, *meta-design*, and *testable hypotheses* (Abbasi and Chen 2008; Walls et al. 1992). Figure 1 depicts our BOLT design framework.

According to Walls et al. (1992), kernel theories are derived from the natural and social sciences and are used to govern meta-requirements. However, as noted by Arazy et al. (2010), theories from those domains are rarely used as-is because their scope and granularity are often inadequate for a specific design problem. Hence, we draw on multiple behavioral studies (Baron and Kenny 1986; Larsen and Bong 2016; Larsen et al. 2019; Weber 2012) as *kernel theories* to identify what constitutes behavioral knowledge. Specifically, behavioral knowledge can be considered as theories encompassing originating, extending, and subscribing behavioral articles (theory instances), and each of these articles is a specialized ontology whose core parts include *constructs* and their *relationships*. Accordingly, the *meta-requirements* frame behavioral knowledge disembedding as a specific ontology learning process (Buitelaar et al. 2005) which needs to support the extraction of *hypotheses (terms)*, *variables (concepts)*, *theoretical relationships (non-taxonomic relations)*, and *synonymous relationships (taxonomic relations)*<sup>2</sup> from behavioral articles. The *meta-design* identifies the underlying tasks and provides viable techniques. Four BOLT tasks are identified as a result: *hypothesis extraction*, *variable extraction*, *theoretical relationship extraction*, and *synonymous relationship identification*. Based on the ontology learning and pertinent NLP literature, we organize viable techniques into two categories—*linguistics* and *statistics/machine learning (ML)*. For each BOLT task, a thoughtful selection and coordination of techniques across these two categories are needed. In this study, we present the best potential techniques

<sup>2</sup>Synonymous relationships depict whether two or more variables are referring to the same underlying meaning, which could be used to build taxonomic relationships. For example, *performance expectancy* in Venkatesh et al. (2003) and *perceived usefulness* in Venkatesh and Morris (2000) are both measuring an individual's perception of a system's usefulness, despite of different words. Hence, they can be considered as hyponyms of the more general construct *perceived usefulness*.



**Figure 1. Behavior Ontology Learning from Text (BOLT) Framework**

(as of the time of publication) as an implementation example. Interested researchers can replace them with better techniques when linguistics and statistics/ML methods advance in the future. Finally, we propose *testable hypotheses* to empirically evaluate how well the proposed meta-design meets the meta-requirements. These hypotheses involve both the ability to extract behavioral knowledge from behavioral articles and the ability to enhance behavioral researchers' information-seeking outcomes. To test them, we propose a multifaceted evaluation solution that includes method and system comparisons, randomized user experiments, and applicability checks. In the following section, we discuss the four components of the BOLT framework in detail.

### Kernel Theories

As noted, there is lack of consensus on how to best define behavioral knowledge and its key components. Weber (2012) argues that theoretical development is a central behavioral

research endeavor. Hence, theories could represent the most important type of behavioral knowledge. Larsen et al. (2019) suggest that a theory consists of a set of publications, including the originating publication, the most influential extensions of the original article, and all theory-subscribing articles. Collectively, these articles are referred to as *theory instances*. According to Weber (p. 3), a theory instance is “a particular kind of model that is intended to account for some subset of phenomena in the real world.”<sup>3</sup> Specifically, the subset of phenomena usually pertains to classes of things in a domain, and the model is an abstracted, simplified, concise representation that explains and predicts a phenomenon (Parsons and Wand 2013). In this light, theory instances “can be conceived as *specialized ontologies*—instances of [Bunge’s (1977,

<sup>3</sup>Weber’s notion of theory is best aligned with Gregor’s (2006) Type IV theory—a theory for explanation and prediction.

1979)<sup>4</sup>] general ontology (a theory about the nature of and makeup of the real world, in general)" (Weber 2012, p. 3).

The core parts of a theory instance include *constructs*, their *relationships*, and the *state* they cover (Weber 2012). In behavioral research, a construct represents "an *attribute in general* of some *class of things* in its domain" (Weber 2012, p. 7). Constructs serve a central role in a theory instance because their definition directly governs the meaning of construct relationships and the state space of a theory. Construct relationships can be in the form of correlation, causation, or synonymous relationships. Correlation or causation, referred to as theoretical relationships hereafter, can be categorized as *main effect*, *moderation*, and *mediation* (Baron and Kenny 1986). Main effect pertains to a direct theoretical relationship between two constructs, moderation involves a third construct affecting the strength or direction of a theoretical relationship, and mediation entails an intermediate construct between two theoretically related constructs. In contrast, a synonymous relationship represents an "is-a" association between different constructs, within or across articles, referring to the same underlying meaning (Larsen and Bong 2016). For example, in Venkatesh and Morris (2000), the two construct mentions of *behavioral intention* in hypotheses H1 and H2a refer to the same meaning; the construct *performance expectancy* in Venkatesh et al. (2003) is synonymous with that of *perceived usefulness* proposed by Davis (1989). Finally, the state of a theory instance is a conceivable state space that falls within a theory's boundary, which is determined by "the range of values that each construct in the theory might cover" (Weber 2012, p. 11). Taken together, this ontology-centric view of behavioral knowledge afforded by the amalgamation of the aforementioned kernel theories provides an appropriate mechanism for disembedding behavioral knowledge.

## Meta-Requirements

As noted, behavioral knowledge could be considered to be comprised of theories encompassing multiple theory instances, each of which can be conceived as a specialized Bunge's (1977, 1979) ontology. The core parts of each theory instance include constructs, their relationships, and the state they cover. Therefore, effective behavioral knowledge disembedding calls for behavioral ontology learning capable of extracting those parts.

Ontology learning pertains to the development and use of various automated techniques to extract the key components

<sup>4</sup>We acknowledge that alternative mapping between behavioral theories and ontologies may exist. In this paper, we choose Weber's mapping to align behavioral knowledge disembedding with the ontology learning layer cake.

of ontology from large-scale textual data (Buitelaar et al. 2005). The goal for ontology learning from text is to bridge the gap in a data context that "scores highest on availability and lowest on accessibility" (Biemann 2005, p. 79)—an objective that nicely parallels our behavioral knowledge inaccessibility alleviation objective. Buitelaar et al. (2005) synthesized the ontology learning literature into a core set of five "layer cake" outputs: *terms*, *concepts*, *taxonomic relations*, *non-taxonomic relations*, and *axioms*. The outputs above are ordered, meaning that each output is a prerequisite for obtaining the next.

- *Terms* are lexical components that contain important pieces for an ontology.
- *Concepts* are formed by leveraging terms to represent objects.
- *Taxonomic and non-taxonomic relations* depict relationships between concepts. *Taxonomic relations* are focused on extracting "is-a" relations (hypernym/hyponym). An example would be "a duck is a type of the concept waterfowl." *Non-taxonomic relations* are non-hierarchical relations. For instance, "a duck is often 'found near' ponds."
- *Axioms* are rules defined over concepts.

In the behavioral ontology learning context, *variables*<sup>5</sup> represent general attributes of some class of things covering a phenomenon and are best aligned with *concepts*. *Synonymous relationships* depict whether two or more variables are referring to the same underlying meaning. These relationships can be used to build a construct hierarchy. Hence, they can be mapped to *taxonomic relations*. *Theoretical relationships* representing correlation and causation are best related to *non-taxonomic relations*. An important starting point in ontology learning, however, is to identify lexical components that encompass variables and relationships. Fortunately, in behavioral research, an article belonging to Weber's notion of theory—theory for explanation and prediction—usually presents the behavioral theory through hypotheses that describe relationships between variables.<sup>6</sup> Hence, we consider *hypotheses* as the *terms* of an ontology learning layer cake.

<sup>5</sup>The term *variable* encompasses behavioral constructs and some non-construct variables that play a key role in the theory (e.g., demographics).

<sup>6</sup>We randomly sampled 40 articles from *MIS Quarterly* and *Journal of Applied Psychology* and had domain experts place the contained hypotheses into two classes: "supported" and "unsupported." The results showed that hypotheses were unsupported 23.7% to 31% of the time in the respective publications. Unsupported hypotheses are important drivers of theoretical progress (Popper 1959). Indeed, in meta-analysis, supported and unsupported relationships are equally important.

In summary, behavioral ontology learning entails four ordered layer cake outputs: *hypotheses (terms)*, *constructs (concepts)*, *theoretical relationships (non-taxonomic relations)*, and *synonymous relationships (taxonomic relations)*. We acknowledge that other behavioral ontology parts, such as the state a theory covers, are also important to disembed (Schryen et al. 2017; Weber 2012); however, we leave these for future research.

## Meta-Design

Based on the aforementioned meta-requirements, the underlying tasks for disembedding behavioral knowledge are *hypothesis extraction*, *variable extraction*, *theoretical relationships extraction*, and *synonymous relationships identification*. A plethora of ontology learning techniques can be leveraged to support these tasks, two of which are pertinent to our context: linguistics and statistics (Wong et al. 2012). *Linguistics* techniques extract linguistic features such as parts of speech (POS), syntactical structure analysis, and dependency analysis. When texts follow a prescribed linguistic pattern, linguistic rules can be derived to extract relevant lexical components for ontology building. *Statistics* techniques are primarily derived from information retrieval, machine learning (ML), and data mining literature. These types of techniques usually rely on linguistic components as underlying input features. Sample techniques include co-occurrence analysis, latent semantic analysis, clustering, and association rules. However, ML approaches, such as classification, sequence labeling techniques, deep learning, and support vector machines (SVMs), that are popular in the state-of-the-art text mining literature for concept and concept relation extraction have been underutilized in the ontology learning domain (Asim et al. 2018; Wong et al. 2012). To better emphasize the immense potential of ML methods such as deep learning, we expand the statistics techniques category in our framework to be *statistics/ML* techniques. In the following section, we describe each BOLT task and identify the best potential supporting techniques from the linguistics and statistics/ML categories. The alternative techniques are described in Appendix A.

## Hypothesis Extraction

The initial task in BOLT is *hypothesis extraction*. Hypothesis formats for several behavioral disciplines (e.g., IS, management, marketing) are formalized. For example, the following shows a typical hypothesis from a behavioral research article by Venkatesh and Morris (2000), which stands as an independent paragraph:

H1: Perceived usefulness will influence behavioral intention to use a system more strongly for men than it will influence women.

A formatting rule could be derived from the above hypothesis: “a capitalized H + a number [+a letter] + a colon + a capitalized word + a string of words + a period.” Similarly, additional rules could be generated by enumerating a sufficient number of articles. Per the ontology learning literature, when the hypothesis text elements have apparent linguistic cues, *rule-based extraction* may be best suited for extracting morphological patterns.

However, hypotheses in many other behavioral research areas may not follow the traditional format identified above. In such situations, *statistics/ML* techniques, such as text classification, can be used to discover appropriate patterns. These methods use training data sets to build text classifiers for automatically predicting class label (i.e., whether a sentence is a hypothesis) based on extensive manual feature engineering. An example baseline method is maximum entropy (Berger et al. 1996). The recent development in ML affords opportunities to further enhance the classification performance (Cho et al. 2014). Examples include *word embedding* (Mikolov et al. 2013) for automated feature engineering and *deep learning* such as convolutional neural networks (CNN) (LeCun et al. 1998) and recurrent neural networks (RNN) (Mikolov et al. 2010) for modeling nonlinear, complex patterns. Word embedding encapsulates word-level distributional semantics. Word2Vec (Mikolov et al. 2013) and GloVe (Pennington et al. 2014) are two common methods for deriving word embeddings. The former uses a neural network model to represent a word’s distributional semantics by examining its surrounding words, and the latter uses a matrix decomposition method over a large-scale data set to model similarities between words. The CNN method utilizes convolutional filters to learn character-level morphological patterns or local features that are critical to differentiating classes of texts (e.g., Kim 2014). The RNN method models long-distance dependency between linguistic components (Hochreiter and Schmidhuber 1997). In summary, due to variances in hypothesis articulation and formatting across authors and disciplines, a hybrid method that combines rule-based techniques and ML approaches is well suited for hypothesis extraction.

## Variable Extraction

The second step in the BOLT framework involves deriving variables from the extracted hypotheses. The following is an example showing variables contained in the first hypothesis of

Venkatesh and Morris (2000). Particularly, perceived usefulness and behavioral intention are constructs, and men and women are the values for the gender variable. We use inside, outside, beginning (IOB) tagging to tag these variables, where “B” marks the beginning word of a variable, “I” labels the words inside a variable, and “O” represents the words outside a variable.

H1/O :/O **Perceived/B usefulness/I** will/O influence/O **behavioral/B intention/I** to/O use/O a/O system/O more/O strongly/O for/O **men/B** than/O it/O will/O for/O **women/B./O**

A survey of the existing techniques addressing concept or entity extraction (suitable for variable extraction) reveals that *statistics/ML* methods are highly applicable (Hobbs and Riloff 2010). Specifically, supervised ML methods convert variable extraction into a sequence labeling problem, where each word in a hypothesis has a specific class label (e.g., the IOB tags), and the dependency among the class labels is explicitly considered (Lafferty et al. 2001). Hence, the sequence labeling task is to predict the most probable class label for each word, depending on its linguistic features as well as its relationship to the surrounding words with their linguistic features. The status quo *statistics/ML* methods for sequence labeling problems are feature-based methods, such as the hidden Markov model (HMM) (Rabiner 1989) and conditional random fields (CRF) (Lafferty et al. 2001), whose performances are heavily dependent upon labor-intensive feature engineering.

The recent advances in ML present opportunities for enhancing performance by involving a hybridized deep learning method (Yadav and Bethard, 2018). Such methods could make use of character-level CNN to incorporate morphological patterns (e.g., capitalization in words) and bidirectional long short-term memory (Bi-LSTM) (Hochreiter and Schmidhuber 1997), a specific type of RNN, to characterize the dependency among linguistic inputs. In addition, incorporating a CRF layer in a deep learning architecture may further enhance performance by considering the dependency among class labels (e.g., Ma and Hovey 2016). Additionally, incorporating word embedding that reflects linguistic richness and domain knowledge could also help (Huang et al. 2015). For example, certain types of words and phrases are known to be more likely to appear in variables (e.g., noun phrases starting with the word “perceived”). This information could be leveraged in customized word embedding of a deep learning architecture, even with limited training data (Huang et al. 2015). In summary, methods at the intersection of customized word embedding, domain-adapted features/lexicons, and hybridized deep learning architectures may be well suited for variable extraction.

## Theoretical Relationship Extraction

Once variables embedded in hypotheses are extracted, the third task involves identifying relationships among them. The following examples show three hypotheses from Krosgaard et al. (2002) that contain main, moderation, and mediation relationships, respectively. Specifically, managerial trustworthy behavior is an antecedent (AT), and employees’ trust in the manager is a consequent (CT), perceived fairness of human resource policies is a moderator (MOD), and employees’ attributions of responsibility for an event are mediators (MED).

**Hypothesis 3 (main):** [Managerial trustworthy behavior]<sub>AT</sub>, in the form of communication and concern, is positively related to [employees’ trust in the manager]<sub>CT</sub>.

**Hypothesis 4 (moderation):** The relationship between [managerial trustworthy behavior]<sub>AT</sub> and [employees’ trust in the manager]<sub>CT</sub> is moderated by the [perceived fairness of human resource policies]<sub>MOD</sub>.

**Hypothesis 5 (mediation):** The relationship between [managerial trustworthy behavior]<sub>AT</sub> and [employees’ trust in the manager]<sub>CT</sub> is mediated by [employees’ attributions of responsibility for the event]<sub>MED</sub>.

To identify such relationships, the syntactic features of the hypothesis are critical. For example, *managerial trustworthy behavior* is the subject and *employees’ trust in the manager* is the object of a verb phrase containing “is positively related to.” Additionally, the moderation and mediation natures of the relationship constitute important behavioral knowledge. Based on several studies (e.g., Maynard et al. 2009; Tan et al. 2016), we posit that effectively extracting complex domain-specific relationships may require multi-stage approaches that combine statistics/ML- and linguistics-based methods. In this vein, SVM (Cortes and Vapnik 1995), motivated by statistical learning theory (Vapnik 1998), has been recognized as a strong performer for relation extraction (Zhou et al. 2005).

## Synonymous Relationship Identification

The final task is to identify synonymous variables within an article or across articles. Linguistics- and statistics/ML-based methods are commonly fused to solve this problem, including lexical similarity analysis, latent semantic analysis, and semantic lexicon. Lexical similarity analysis assesses the degree of similarity between two texts at the lexical level and is a common ontology learning method (Gefen et al. 2017; Sharman et al. 2007). Typical methods include minimum edit

distance, which measures the minimum number of single-character edits (insertions, deletions, or substitutions) to convert one text string into another (Strube et al. 2002). Semantic lexicon is a popular resource for ontology learning (Wong et al. 2012) that consists of predefined concepts and relations and can be used for identifying terms, concepts, taxonomic, and non-taxonomic relations. Well-known semantic lexicons include WordNet (Fellbaum 1998; Miller 1995) and the Unified Medical Language System (Bodenreider 2004). To identify synonymous variables within an article, lexical similarity analysis and semantic lexicon could be used due to consistent naming conventions enforced in academic articles. For variables across articles, one state-of-the-art technique is the construct similarity algorithm proposed by Larsen and Bong (2016), which uses multiple semantic lexicons (e.g., WordNet) and a customized latent semantic analysis fused with other lexical similarity measures to extract the hypernym/hyponym relationships among constructs.

### Testable Hypotheses

Testable hypotheses are intended to evaluate how well the proposed meta-design satisfies our meta-requirements (Walls et al. 1992). For the proposed design framework, this entails two aspects: the ability to extract behavioral knowledge from behavioral articles and the ability to enhance information-seeking outcomes. A multifaceted evaluation solution is needed to address these two aspects (Gill and Hevner 2013; Hevner et al. 2004). Specifically, the former hypothesis requires rigorous comparisons with alternative ontology learning techniques and systems, and the latter calls for user experiments to shed quantitative and qualitative light on when, to whom, how, and to what extent the proposed meta-design enhances behavioral researchers' information-seeking outcomes (Abbasi et al. 2018). However, all of the evaluation methods require instantiations of the proposed design framework as a basis to evaluate the effectiveness and applicability of our meta-design (Abbasi and Chen 2008).

The remainder of the paper is organized as follows. We first describe the TheoryOn systems (developed as an instantiation of the proposed BOLT design framework). The two ensuing sections provide experimental evaluations of the TheoryOn system and its underlying framework with regards to two testable hypotheses—the ability to extract behavioral knowledge and the ability to enhance information-seeking outcomes. The discussion of empirical insights and generalizability of the proposed design framework and instantiation are offered. We conclude with a summary of our research contributions and potential future directions.

## TheoryOn: An Instantiation of the Proposed Design Framework

Using the design guidelines prescribed by the BOLT framework, Figure 2 depicts the system diagram for the TheoryOn instantiation. Each article underwent *hypothesis extraction*, *variable extraction*, and *relationship extraction* using the techniques prescribed by our framework. Specifically, hypothesis extraction used a sentence classifier combining deep learning and rule-based extraction. For variable and relation extraction, we developed an SVM classifier with a deep learning-informed multi-stage tree composite kernel (DLMTCK). With the extracted variables and hypotheses, the theoretical network was assembled in the *theoretical network construction* step and was indexed and placed inside the TheoryOn *Search & Visualization* application. Both steps are instantiations for the *Synonymous Relationship Identification* task in BOLT. In following design guidelines of BOLT as closely as possible, we assured *instantiation validity* of TheoryOn (Lukyanenko et al. 2015). Next, we discuss each step of TheoryOn in detail.

### Hypothesis Extraction

Following the guidelines offered by the BOLT framework, TheoryOn utilizes a sentence classifier that couples rule-based and deep learning methods to extract hypotheses. We first used a rule-based approach to identify the hypothesis extraction rules. The details of the rule identification process are depicted in Appendix B. These extraction rules were then coupled with a deep learning method, depicted in Figure 3 and described below.

The word2vec method was used to create pre-trained word embeddings from article texts in order to efficiently incorporate distributional semantic information. This word embedding was then inputted into a Bi-LSTM to learn the long-distance dependencies among words. Additionally, we represented rule-based features as one-hot vectors, indicating whether the sentence contained an extraction pattern in Appendix B. We also incorporated the sentence order features, motivated by the fact that hypotheses often appear earlier in articles. All these features were concatenated and put into a dense layer to calculate the final classification probabilities. Later, in the evaluation section, we justify our design choices guided by BOLT by benchmarking our hybrid classifier against standard deep learning for text classification, a rule-based approach, and several feature-based methods.

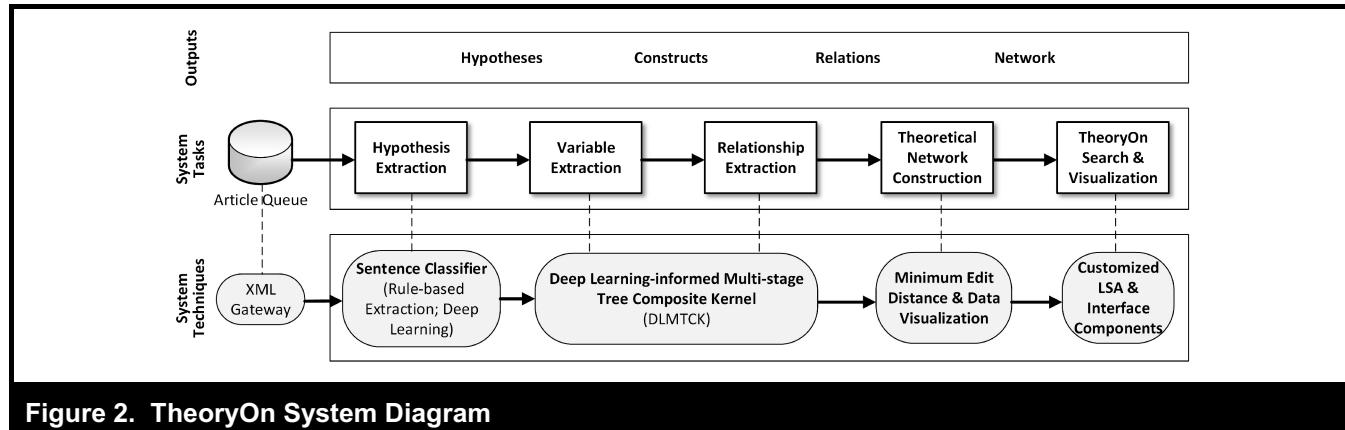


Figure 2. TheoryOn System Diagram

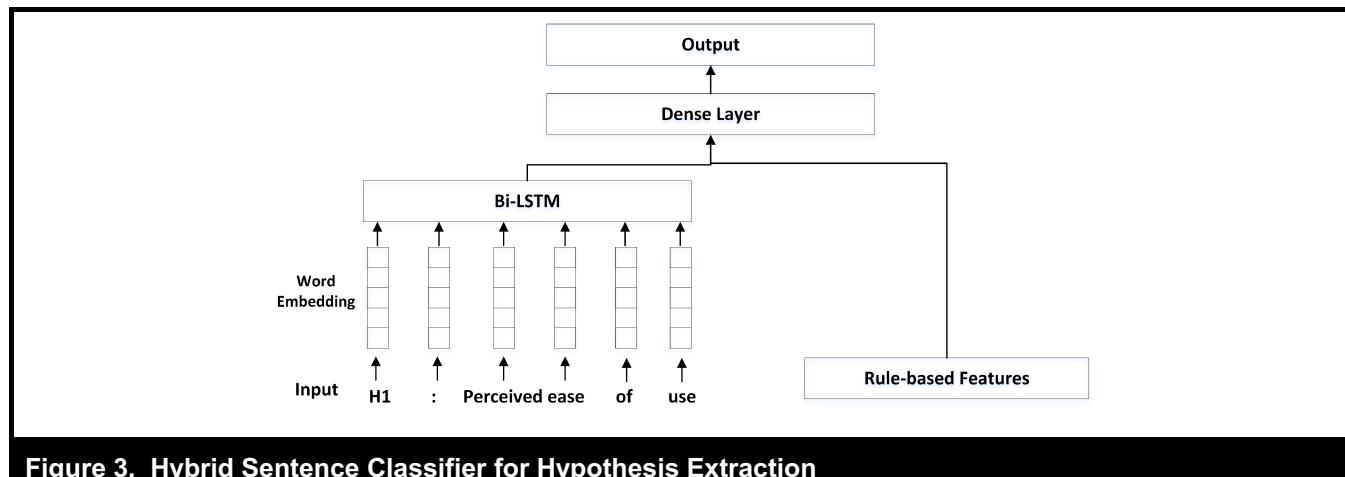


Figure 3. Hybrid Sentence Classifier for Hypothesis Extraction

### Variable and Theoretical Relationship Extraction

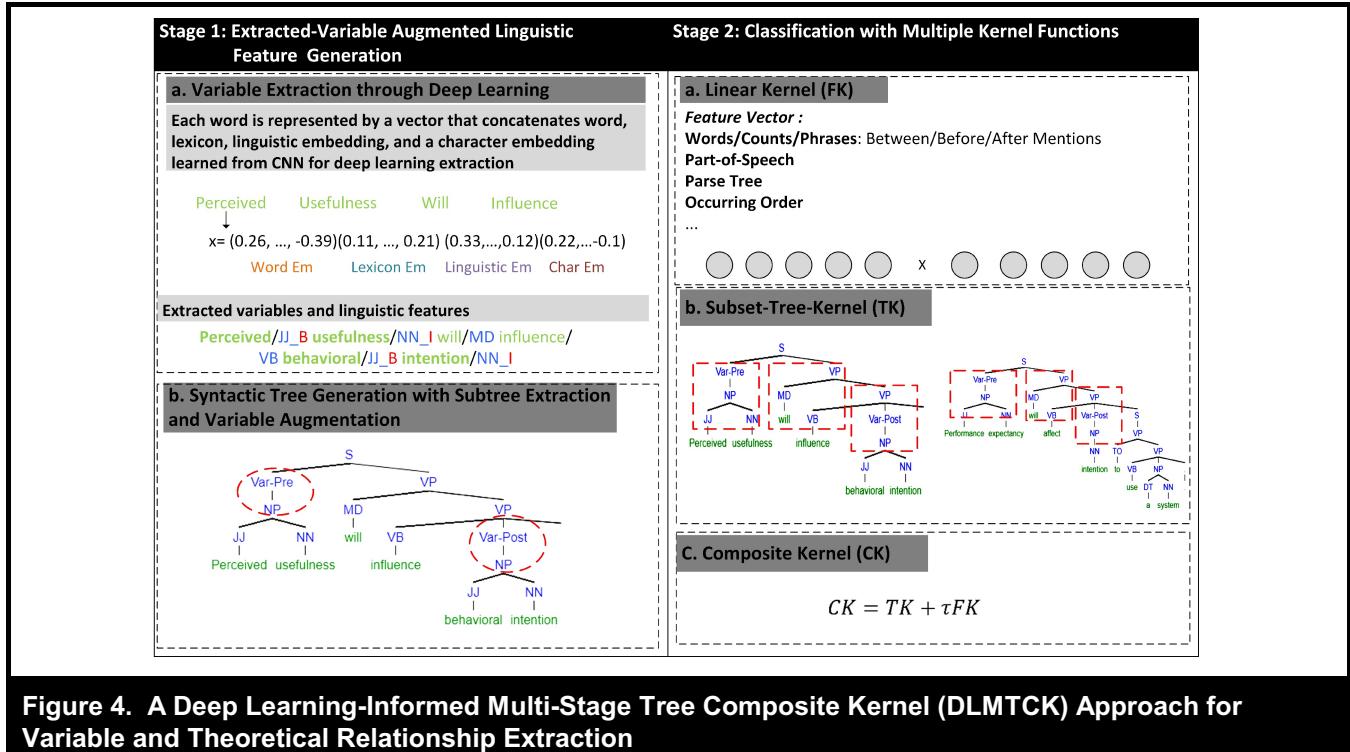
According to the BOLT framework, the performance of theoretical relationship extraction was heavily dependent on the accuracy of variable extraction. Therefore, we combined these two steps to allow for fast iterations in model tuning. Accordingly, we propose a two-stage labeled tree kernel, with the first stage focusing on extracting rich linguistic patterns that are augmented by deep learning-informed variable extraction and the second stage encompassing an SVM that fuses the domain-specific linguistic patterns in a composite kernel function. Our proposed DLMTCK approach is illustrated in Figure 4.

#### Stage 1: Extracted-Variable Augmented Linguistic Feature Generation

The first step in Stage 1 involved extracting variables from the hypotheses using deep learning (Stage 1(a) of Figure 4).

Guided by BOLT, the architecture included a character-level CNN, word embedding, lexicon embedding, and linguistic embedding. Character-level CNN modeled the morphological patterns for each word. After the CNN layer, each character was represented as a fixed dimensional vector. Max pooling was applied to aggregate the character-level embedding to the word level, which was then fed into the Bi-LSTM layer. Pre-trained word embedding was leveraged to represent word-level distributional semantics. The semantic lexicon and linguistic embedding enriched domain adaptation and learned syntactic patterns specific to behavioral knowledge. Specifically, lexical embedding leveraged a one-hot vector to represent whether a word was contained in a behavioral lexicon. The behavioral lexicon contains representative variable words derived from training data and domain experts. The linguistics features embedding represents relevant linguistic features and their interactions such as POS and chunk as multi-hot vectors.<sup>7</sup> These vectors were fed into an

<sup>7</sup><https://www.clips.uantwerpen.be/pages/mbsp-tags>



**Figure 4. A Deep Learning-Informed Multi-Stage Tree Composite Kernel (DLMTCK) Approach for Variable and Theoretical Relationship Extraction**

embedding layer to generate a dense vector representation, which was concatenated with the pre-trained word embedding and fed into a Bi-LSTM layer, followed by a CRF layer to model the dependencies among IOB tags. The detailed deep learning architecture is presented in Figure 5.

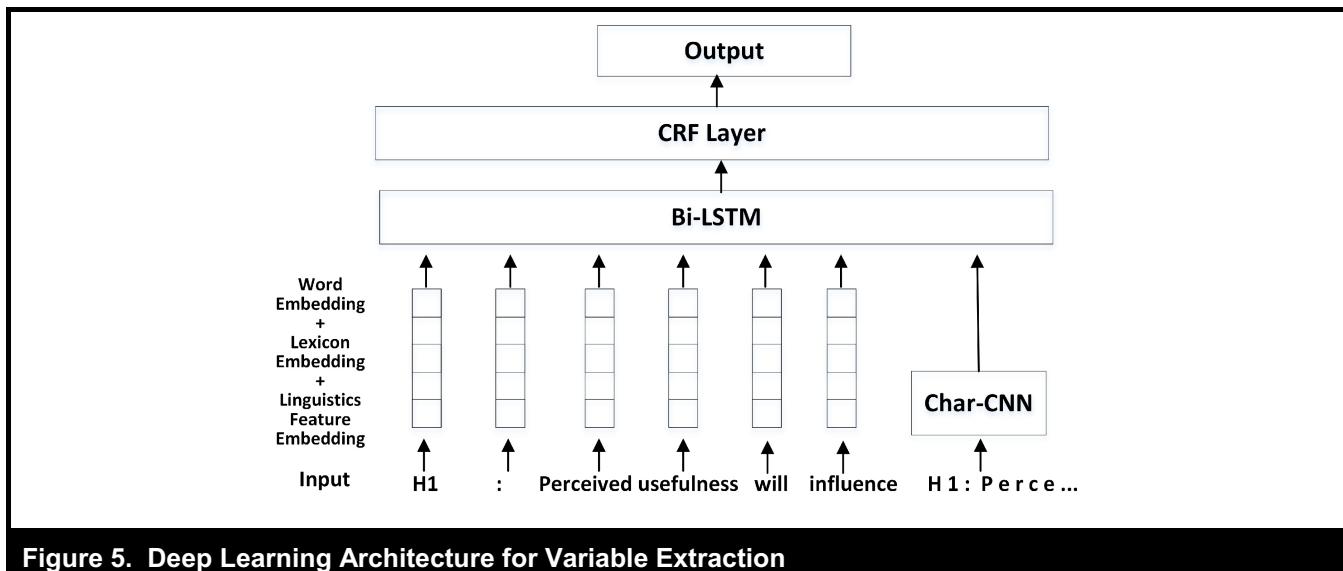
Upon completion of variable extraction, in Stage 1b, we derived a syntactic tree for each hypothesis with subtree extraction and variable augmentation. Specifically, for any two variables in a hypothesis, we first extracted all subtrees that encompassed these two variables and then enriched the subtree with variable indicators. For example, in Figure 4, Stage 1(b), Var-Pre indicates the minimal phrase that contained the first variable. These rich subtree patterns were subsequently utilized in Stage 2, which is discussed later.

## Stage 2: Classification With Multiple Kernel Functions

In the second stage, a composite kernel SVM was used to predict the theoretical relationships between variables. SVM uses the maximum margin principle to find two parallel hyperplanes that can divide a set of data points into two classes (e.g., having a particular relationship or not), in which the margin is defined by the perpendicular distance between these two parallel hyperplanes (Cristianini and Shawe-Taylor

2000). The hope is that the larger the margin, the smaller the generalization error. For our relation extraction problem, this is translated into finding the optimal hyperplanes for three relationship types: the main effect (variable1, variable2), moderation (variable1, moderator, variable2), and mediation (variable1, mediator, variable2). However, moderation and mediation involve three variables, whereas SVM relation extraction typically concerns the classification of relationships involving two variables. We therefore decomposed a ternary relationship for moderation into two binary relationships of moderation (moderator, variable) and one binary relationship of the main effect (variable1, variable2), and a ternary relationship for mediation into two binary relationships of mediation (mediator, variable) and one binary relationship of the main effect (variable1, variable2).

Consequently, for moderation and mediation relationships, we first classified the derived binary relationships and assembled them into ternary relationships. For each derived binary relationship, we used a composite kernel function to reflect the diverse linguistics patterns. In SVM, a kernel function measures the similarity between two feature vectors by mapping them to a higher-dimensional space so that an optimal hyperplane could be found. It can also be tailored to incorporate domain-specific knowledge (Burges 1998; Muller et al. 2001). Composite kernels are well suited to incorporate broad, relevant features while reducing the risk of over-fitting



**Figure 5. Deep Learning Architecture for Variable Extraction**

(Collins and Duffy 2002; Kitchens et al. 2018; Szafranski et al. 2010). Specifically, our composite kernel function (Stage 2c) is a linear combination of two kernel functions (Zhou et al. 2010), including a linear kernel function (Stage 2a) characterizing the flat linguistic patterns defined by Zhou et al. (2005) and a sub-tree kernel function (Stage 2b) incorporating the variable augmented subtree features from the first stage (Collins and Duffy 2002). A detailed description of these kernel functions is shown in Appendix C.

Once the relationship type of any variable pair in a hypothesis was determined, we consolidated the binary relationships into ternary relationships based on shared constructs. The three types can represent hypotheses with any number of variables. The unit of analysis for evaluation is thus based on how many of these types is extracted correctly.

### Theoretical Network Construction

TheoryOn visualizes an article's theoretical network by grouping the variables shared across its hypotheses. For example, after variable and relation extraction, H1 and H2a in Venkatesh and Morris (2000) can be represented as solid boxes and arrows in Figure 6(2), in which men and women are the values of a gender variable moderating the relationship between perceived usefulness and behavioral intention, as well as that between perceived ease of use and behavioral intention.

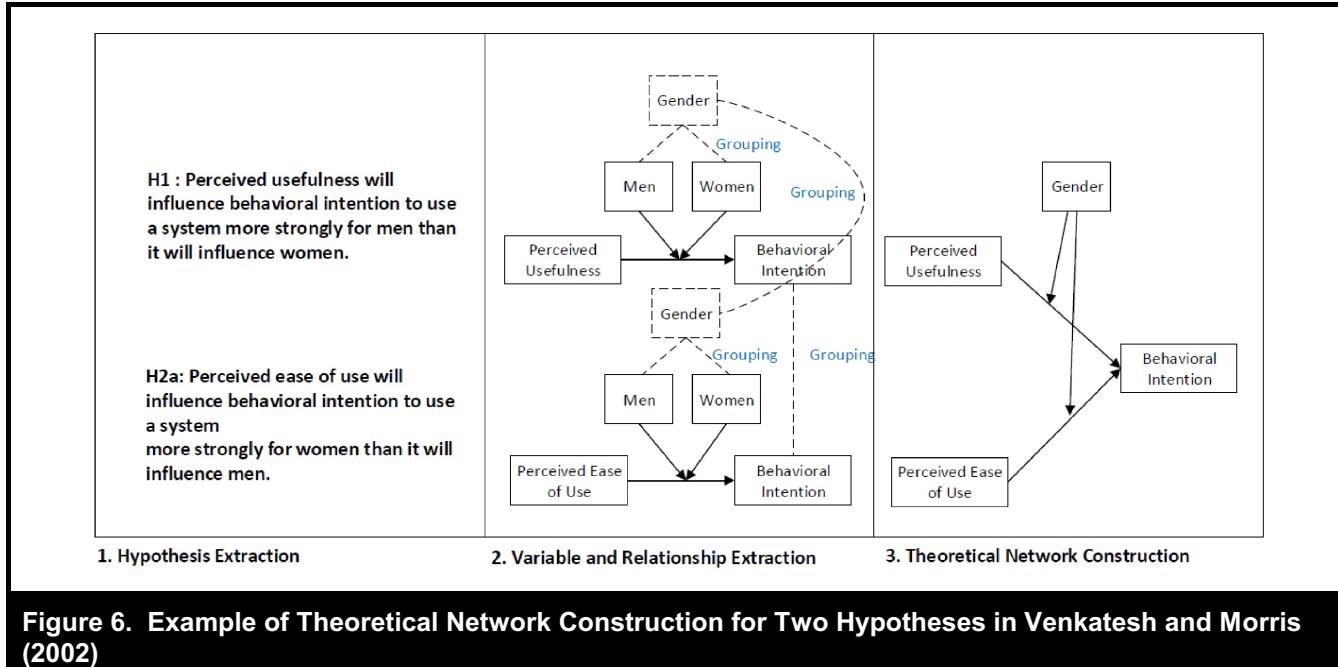
In order to create a succinct theoretical network visualization, "men" and "women" are grouped as a "gender" variable

through a semantic lexicon (e.g., women, men, boy, and girl are hyponyms for gender). Furthermore, two "behavioral intention" and two "gender" variables from H1 and H2a, respectively, are grouped together using lexical similarity analysis. Specifically, a minimum edit distance measure was used to calculate their similarity, with the threshold determined empirically through validation. Combining all hypotheses through shared variables could automatically construct a theoretical network for each behavioral article.

### TheoryOn Search and Visualization

Automatically extracting theoretical networks allows users to conduct an ontology-centric search, in which a user types a variable as a search query to obtain a list of relevant theoretical networks. To accomplish this objective, synonymous relationships among variables from different articles should be identified. For example, the construct *performance expectancy* in Venkatesh et al. (2003) is synonymous with the construct *perceived usefulness* in Venkatesh and Morris despite comprising different words. When a user types the search query "perceived usefulness," both articles should be returned.

Following the guidance of BOLT, we utilized customized latent semantic analytics (LSA) (Larsen and Bong 2016) coupled with a standard Lucene keyword search algorithm (McCandless et al. 2010). TheoryOn users are given the option of selecting a keyword search or a combination of keyword and LSA search, serving multiple stages of the information-seeking process.



**Figure 6. Example of Theoretical Network Construction for Two Hypotheses in Venkatesh and Morris (2002)**

### System Interface

Basing on common behavioral knowledge disembedding use cases related to the key output of the BOLT framework, we developed the following four system functionalities:

- (1) **Construct Search.** TheoryOn allows users to specify a construct in a search query in order to search articles that contain this construct or its synonymous constructs. Figure 7 shows a search for “perceived usefulness” using a combination of keyword and customized LSA search. Each retrieved construct is shown with the theoretical network it belongs to, with the target construct marked in yellow. For details, watch the video “[TheoryOn: Synonymous Construct Search](#).<sup>8</sup>
- (2) **Construct Pair Search.** TheoryOn allows users to specify a construct pair in a search query and to find articles containing those two constructs (see Figure 8). The constructs (marked in yellow) 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](#).<sup>8</sup>

(3) **Theoretically Related Construct Search.** This functionality allows inspection of the theoretical models containing a construct of interest (under “Antecedents” and “Consequents” section), as well as the examination of its antecedents and consequents in a list or plot view (Figure 9). TheoryOn takes the first  $n$  articles returned by the construct search and displays the antecedents to the construct searched for. It then does the same for the consequents. For more details, watch the video “[TheoryOn: Theoretically Related Construct Search](#).<sup>8</sup>

(4) **Theory Integration.** All the related theories can be saved in the left panel and visualized on the canvas (see Figure 10). A user can then integrate theories by clustering synonymous constructs, or the user can customize the theoretical networks by editing, deleting, or adding any nodes and links. For more details, watch the video “[TheoryOn: Theory Integration](#).<sup>8</sup>

### Evaluation: Experiments to Examine Behavioral Knowledge Extraction Performance

To demonstrate the effectiveness of the BOLT framework’s design guidelines, as well as the proposed methods for hypothesis, variable, and relation extraction, we benchmarked our framework-guided techniques with alternative methods. For each task, the labeled set was divided into a training (60%),

<sup>8</sup>Video links lead to an anonymized YouTube channel, compliant with *MIS Quarterly*’s blind review policies. A prototype version of the tool is available at TheoryOn.org.

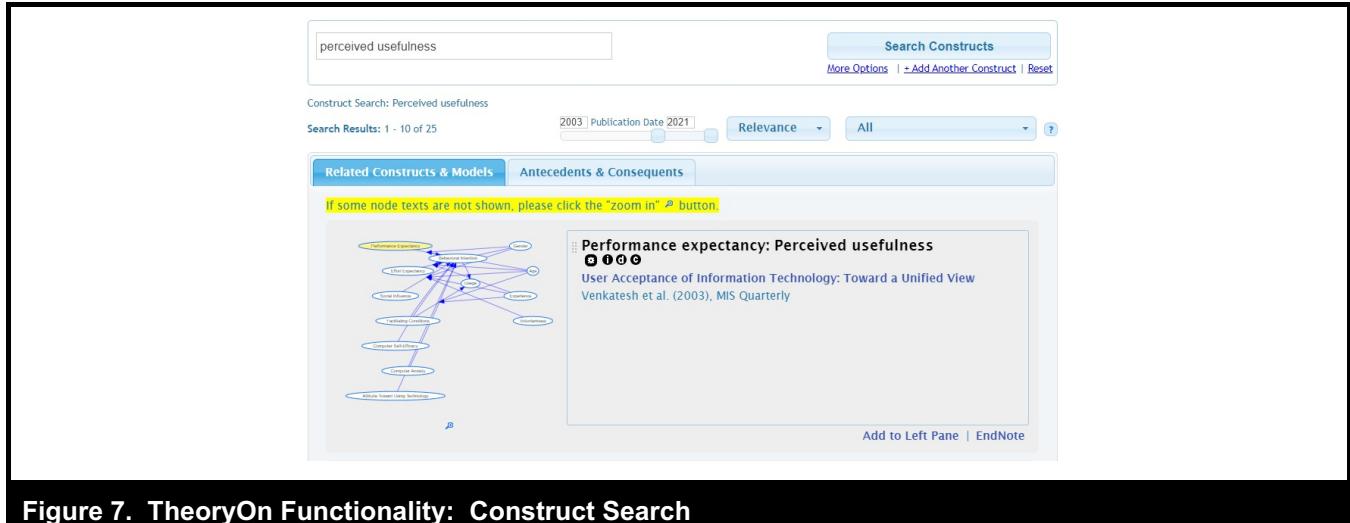


Figure 7. TheoryOn Functionality: Construct Search

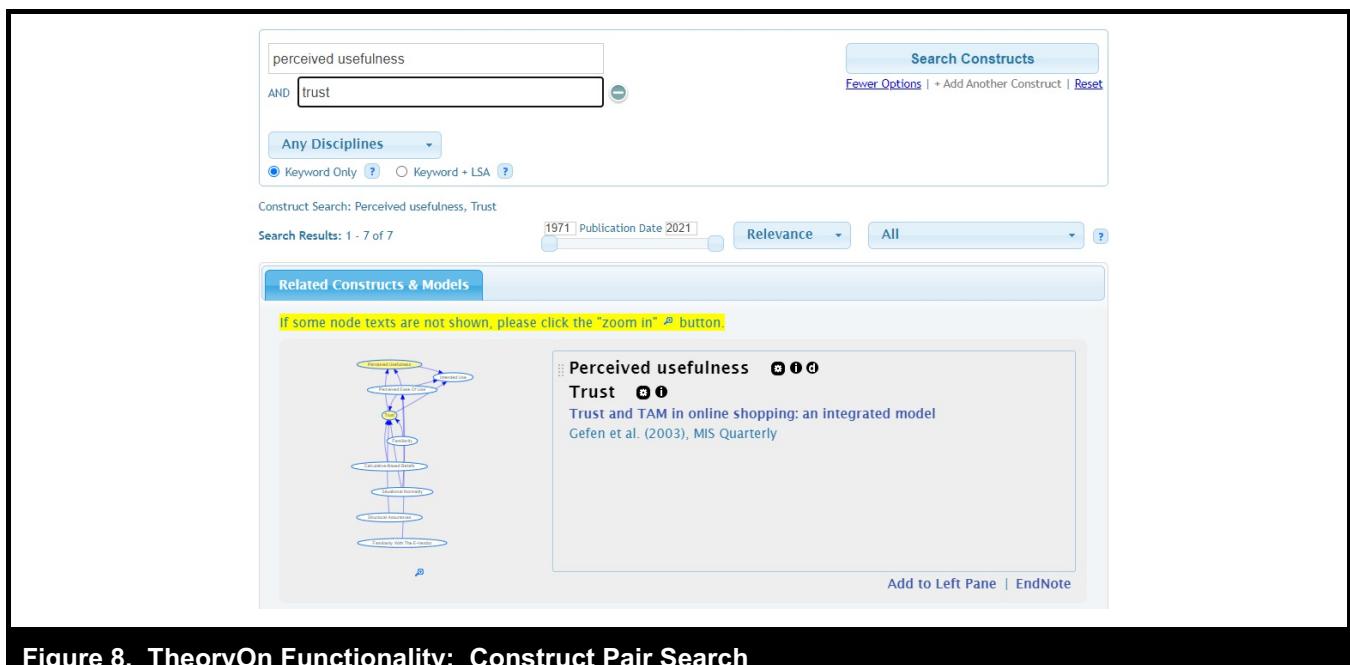


Figure 8. TheoryOn Functionality: Construct Pair Search

a development (20%), and a test set (20%). The training set was used to build the model, the development set was used to search for the optimal learning parameters, and the test set was used to report the model performance. The labeled data consisted of 69 articles from *MIS Quarterly*, 72 articles from *Information Systems Research*, and 145 articles from the *Journal of Applied Psychology* from 1980 to 2009. These three journals were chosen to illustrate the generalizability of our methods across multiple disciplines in behavioral research. Each article was labeled by two annotators with a combined 20+ years of research experience in behavioral

research. The inter-rater reliability, measured by Cohen's kappa, for hypothesis, variable, and relationship extraction was 0.98, 0.75, and 0.82, respectively, indicating agreement levels that are substantial and close to being almost perfect (Landis and Koch 1977). Any disagreement between the two annotators was extensively discussed and resolved, resulting in 1,913 manually extracted hypotheses, 6,020 variable instances, and 3,135 basic relationships (binary or ternary). Two sets of experiments were performed: method comparison and system comparison. In the method comparison, we evaluated our hypotheses, variable, and relationship extraction

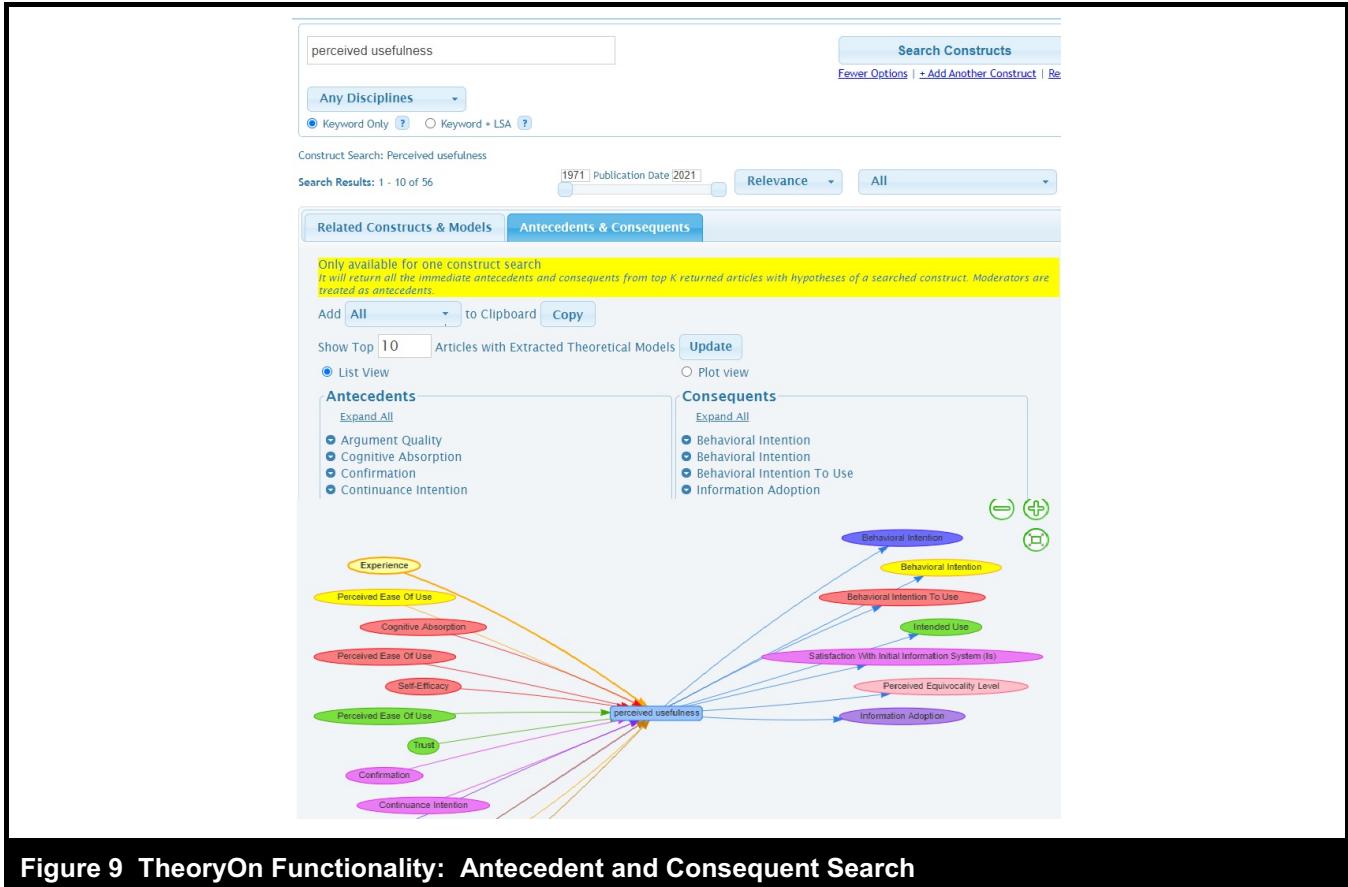


Figure 9 TheoryOn Functionality: Antecedent and Consequent Search

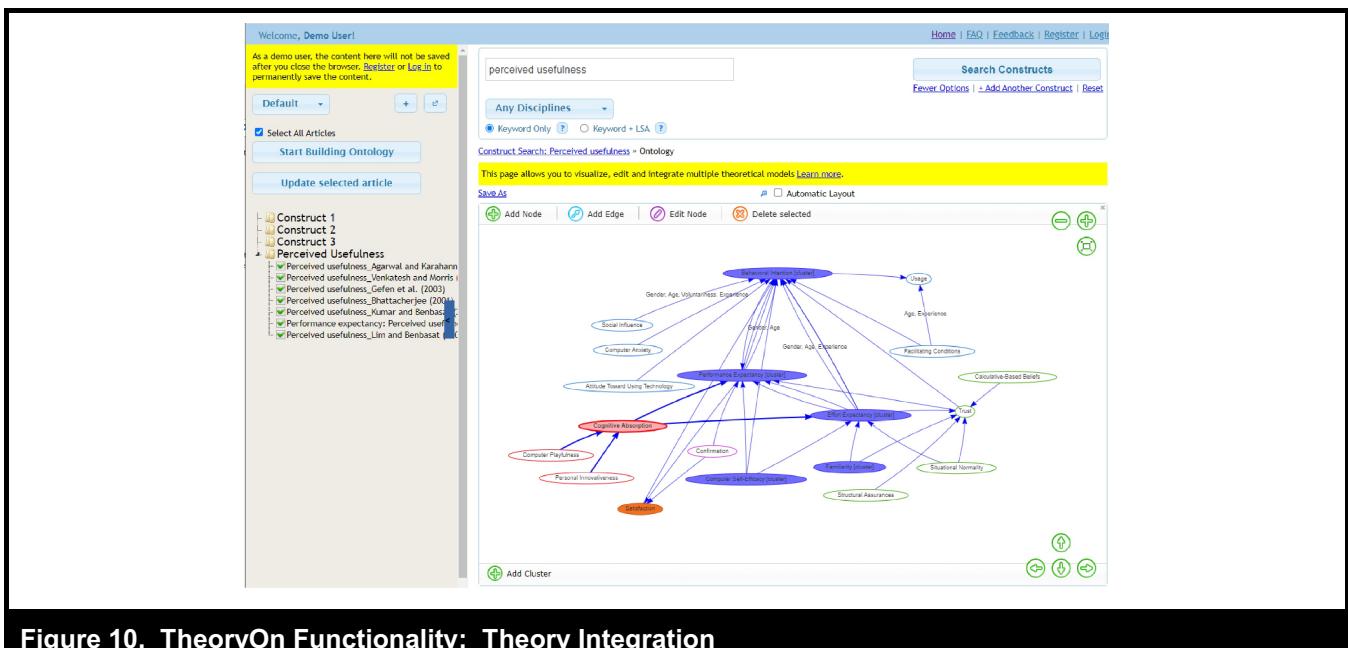


Figure 10. TheoryOn Functionality: Theory Integration

methods against existing techniques. For the system comparison, TheoryOn's variable and relation extraction capabilities were compared with state-of-the-art text ontology learning systems. The details regarding the experiments are given next.

### **Method Comparison Experiments and Results**

For the hypothesis extraction task, we compared the aforementioned hybrid sentence classifier with a rule-based method, feature-based classifiers, and deep learning classifiers. The feature-based classifiers included maximum entropy. The deep learning classifiers included a CNN classifier proposed by Kim (2014), an LSTM classifier by Tang et al. (2015), and a hybrid a BiLSTM-CNN classifier by Zhou et al. (2016). For variable and relation extraction, our proposed DLMTCK method was evaluated against state-of-the-art techniques. The variable extraction benchmarks included the BiLSTM + CRF classifier by Huang et al. (2015), the character-level CNN (CharCNN) + BiLSTM + CRF classifier by Ma and Hovy (2016), CRF, the domain relevance measure (DRM) (Jiang and Tan 2010), the C/NC value (Drymonas et al. 2010), noun phrase term frequency-inverse document frequency (npTFIDF) (Maedche and Staab 2000), and lexicon-driven concept identification through binary predicate (BPLex) (Oliveira et al. 2001). The classifier BiLSTM + CRF inputs a pre-trained word embedding to a BiLSTM layer and classifies IOB tags using a CRF decoder. The CharCNN + BiLSTM + CRF method extends this architecture by adding character-level embedding and a CNN to model the morphological patterns of each word. The DRM method identifies noun phrases and uses domain-specific lexicon coupled with a likelihood test measure to determine phrases that might constitute potential concepts. The C/NC value uses a heuristic measure to score candidate noun phrases based on co-occurrence patterns indicative of concept mentions. The npTFIDF extracts all noun phrases, removes articles and certain descriptive adjectives such as "several" and "many," and computes *tfidf* to eliminate irrelevant terms (i.e., those below a specified threshold). Finally, BPLex derives noun phrase patterns through a binary predicate function that uses part-of-speech tags and syntactic structures.

For relation extraction, the benchmarks included a linear SVM, the verb rule method (Jiang and Tan 2010), association rule mining (Drymonas et al. 2010; Maedche and Staab 2000), and customized semantic template invoking lexicosyntactic patterns (LexSynPatt) (Oliveira et al. 2001; Vargas-Vera et al. 2001). The verb rule method utilized a predefined noun-verb-noun regular expressions that were capable of identifying non-taxonomic relations between constructs.

Association rule mining was used to obtain noun-verb-noun rule sets encompassing antecedent constructs and consequent constructs with appropriate level of support and confidence levels. The LexSynPatt represented relation instances in the training set as item sets encompassing constructs, lexicosyntactic patterns such as verb-based binary predicates.

The experiment results are presented in Table 1. In terms of hypothesis extraction, the hybrid classifier performed better with precision, recall, and  $F_1$ -measure compared with the deep learning methods. The rule-based method had high precision and relatively low recall because of additional patterns residing in the test data. The feature-based methods, such as maximum entropy, performed worse than the deep learning methods.

For variable extraction, the proposed DLMTCK method offered much better performance than the comparison methods did. The performances of CRF and deep learning methods are complementary in the sense that CRF could model class label dependency, whereas deep learning methods could effectively represent input features. HMM was hampered by its reliance on feature token representations and its inability to consider long-distance interdependencies. Concept extraction methods from prior ontology learning studies are designed for extracting general-purpose concepts, which may include valid concepts that are not behavioral constructs or noun phrases that are not exactly matching behavioral construct phrases. Nonetheless, these methods' over-reliance on general noun phrase extraction principles may not be suitable for behavioral ontology learning context.

For relation extraction, DLMTCK outperformed the linear SVM by about four percentage points on F-measure, demonstrating the value of the tree structure approach for the enhanced identification of construct relations. Once again, existing text ontology learning methods were designed for general-purpose relation extraction, which could include relationships that are outside of the theoretical construct relationships or miss relationships that are not connected by verbs. Hence, they could not precisely and comprehensively represent the myriad relation patterns embodied in behavioral texts. Collectively, the results show the efficacy of the meta-design provided by the BOLT framework and demonstrate the utility of the proposed DLMTCK method for variable and relation extraction.

### **System Comparison Experiments and Results**

To examine its system-level performance, TheoryOn was compared with existing text ontology learning systems. To

**Table 1. Method Comparison Results for Hypothesis, Variable, and Relationship Extraction**

		Precision	Recall	F1
<b>Hypothesis Extraction</b>	<i>Hybrid</i>	<b>96.27%</b>	94.26%	<b>95.25%</b>
	BiLSTM-CNN	87.01%	92.69%	89.76%
	BiLSTM	94.04%	90.60%	92.29%
	CNN	80.26%	<b>95.56%</b>	87.25%
	Rule based	93.92%	88.77%	91.28%
	Maximum Entropy	95.12%	81.46%	87.76%
<b>Variable Extraction</b>	<i>DLMTCK</i>	<b>77.07%</b>	<b>76.17%</b>	<b>76.61%</b>
	CNN + BiLSTM + CRF	74.89%	72.58%	73.72%
	BiLSTM + CRF	74.04%	72.50%	73.26%
	CRF	74.43%	70.58%	72.46%
	HMM	60.45%	55.42%	57.83%
	DRM	27.91%	48.92%	35.54%
	C/NC Value	24.44%	45.50%	31.80%
	npTFIDF	25.77%	45.75%	32.97%
	BPLex	23.81%	44.33%	30.98%
<b>Relation Extraction</b>	<i>DLMTCK</i>	<b>88.44%</b>	<b>80.98%</b>	<b>84.54%</b>
	Linear SVM	83.61%	78.04%	80.73%
	Verb Rule Method	63.24%	48.24%	54.73%
	Association Rules	65.33%	48.04%	55.37%
	LexSynPatt	72.29%	56.27%	63.29%

**Table 2. System Comparison Results for Variable and Relationship Extraction**

		Precision	Recall	F1
<b>Variable Extraction</b>	TheoryOn	<b>71.34%</b>	<b>70.33%</b>	<b>70.84%</b>
	CRCTOL	25.15%	43.17%	31.78%
	OntoGain	20.89%	39.83%	27.41%
	TextOnto	22.95%	39.92%	29.15%
	TextStorm	21.10%	38.42%	27.24%
<b>Relation Extraction</b>	TheoryOn	<b>74.05%</b>	<b>64.90%</b>	<b>69.17%</b>
	CRCTOL	38.44%	24.12%	29.64%
	OntoGain	34.85%	22.55%	27.38%
	TextOnto	35.15%	22.75%	27.62%
	TextStorm	31.54%	24.12%	27.33%

select the most appropriate baseline systems, we used the evaluation guidelines proposed by Park et al. (2007), namely, general, extraction, and quality features, as inclusion criteria. Many systems we surveyed were not applicable because they lack sufficient extraction features such as extraction levels and degrees of automation (Park et al. 2007). For example, the FFCA system (Quan et al. 2004) and ASIUM (Faure and Poibeau 2000) do not tackle non-taxonomic relations, and DODDLE-OWL (Morita et al. 2006) uses a semi-automatic extraction process. Consequently, concept–relation–concept tuple-based ontology learning (CRCTOL; Jiang and Tan

2010), OntoGain (Drymonas et al. 2010), Text-To-Onto (Maedche and Staab 2000), and TextStorm (Oliveira et al. 2001) were selected. Because these systems are not designed for behavioral ontology learning and may include other ontology extraction steps, we only selected their relevant components without modification for comparison. Specifically, CRCTOL automatically mines concepts and relations using DRM-based noun phrase extraction and predefined noun-verb-noun patterns. OntoGain uses a C/NC value-based noun phrase extraction algorithm coupled with an association rule mining-based relation extraction method. Text-To-Onto com-

bines syntactic patterns of noun phrases with association rule mining, and TextStorm uses binary predicates for concept and relation extraction. Unlike the method evaluation, the system comparison examined the performance of the ontology learning pipelines, including the error/performance interaction effects between stages. As non-BOLT systems do not have formal hypothesis extraction mechanisms, we began with the extracted hypotheses and focused on the variable and relation extraction stages of the system pipelines.

The results are shown in Table 2. As expected, system pipeline performance was lower relative to the isolated testbed method results depicted in Table 1 because of error propagation. Consistent with the method experiments, TheoryOn offered better recall and  $F_1$ -measures for variable and relation extraction relative to the four comparison text ontology systems. The performance of generic text ontology systems confirms our initial belief that instantiations grounded in BOLT are necessary to make behavioral knowledge disembedding feasible. Next, we performed a user experiment and an applicability check to empirically demonstrate the practical downstream value of TheoryOn's hypothesis, variable, and relation extraction capabilities, which is discussed in the next section.

## Evaluation: User Experiments to Examine Information-Seeking Outcomes

We conducted two user studies, namely a randomized user experiment and an applicability check, to evaluate TheoryOn's ability to enhance behavioral researchers' information-seeking outcomes quantitatively and qualitatively. Specifically, the randomized user experiment compares researchers' performance across four information-seeking tasks among TheoryOn, Google Scholar, and EBSCOhost. The applicability check uses the nominal group technique (NGT) to collect qualitative feedback from behavioral researchers in terms of when, to whom, and how TheoryOn might be beneficial to information-seeking.

### Randomized User Experiment

We selected two full-text search engines, Google Scholar and the Business Source Complete database powered by EBSCOhost, as the benchmarking full-text search engines. Both of them represented, at the time of the experiment, the longest uninterrupted period of full-text coverage for *MIS Quarterly*, *Information Systems Research*, and the *Journal of Applied Psychology* (1990–2009). A total of 52 information systems and organizational behavior Ph.D. students from pro-

grams around the globe were randomly assigned to one of the three experimental groups (TheoryOn, EBSCOhost, or Google Scholar).

We designed the following four tasks for each participant to complete: *synonymous construct search*, *construct pair search*, *antecedent and consequent search*, and *theory integration*, each of which is a common scholarly information task for behavioral research. All four tasks were related to one theory, the technology acceptance model (TAM), to demonstrate a natural progression of knowledge acquisition, curation, and integration in an information-seeking process. TAM was selected because of high awareness, which set up a context in which users of Google Scholar and EBSCOhost were given every opportunity to perform at their peak.

The gold standard for each task was rigorously constructed by a team of two experienced faculty researchers, three doctoral students, and four senior research assistants (research assistants had at least 500 hours of experience in construct extraction from behavioral articles). Following Hevner et al. (2004) and Gill and Hevner (2013), we evaluated TheoryOn's performance using both objective and perceptual evaluations. The objective evaluation compared the construct, article, and theory retrieval performance, including precision and recall (Salton 1989), whereas the subjective evaluation examined the perceived utility of the artifact, reflected by the *perceived usefulness*, *perceived ease of use* and *behavioral intention* constructs from Davis (1989) and Venkatesh et al. (2003). The detailed experiment information regarding the randomization checks, task description, and evaluation procedure are depicted in Appendix D.

### Construct and Theory Retrieval Performance

The results in Table 3 showed that the participants using TheoryOn attained F-measures that were 37% to 121% higher for all tasks, relative to the EBSCOhost and Google Scholar full-text search engines. Specifically, TheoryOn performed especially well in complex tasks, such as antecedent and consequent search, as well as in theory integration. These results demonstrate the viability of TheoryOn for potentially mitigating the knowledge inaccessibility problem that manifests during the scholarly information-seeking process. Compared with EBSCOhost and Google Scholar, TheoryOn could reduce false negatives in search results by up to 158%. This is because TheoryOn directly extracts hypotheses, constructs, and relationships and visualizes them in a user-friendly format, saving researchers precious time and effort otherwise expended extracting and processing behavioral knowledge from articles. The bandwidth freed up by TheoryOn's automated assistance allows users to process more articles (reducing

**Table 3. Percentage Retrieval Performance by Task**

Task	TheoryOn (n = 18)			EBSCOhost (n = 17)			Google Scholar (n = 17)		
	Precision	Recall	F1	Precision	Recall	F1	Precision	Recall	F1
1. Synonymous Construct Search	95.2	27.3	<b>40.1</b>	81.7	16.2	26.4	88.9	18.0	28.5
2. Construct Pair Search	76.7	43.9	<b>51.6</b>	72.0 <sup>+</sup>	24.7	34.9	51.2	34.7	37.2
3a. Antecedent Search	86.3	29.3	<b>41.5</b>	72.2	13.4	21.8	71.6	12.6	20.1
3b. Consequent Search	80.2	23.8	<b>34.7</b>	68.9	16.4	25.3	82.3 <sup>+</sup>	15.8	24.6
4. Theory Integration	77.4	25.4	<b>34.6</b>	61.9	16.0	23.9	62.9	9.8	15.6

Note: <sup>+</sup>not significantly different from TheoryOn ( $p > 0.05$ ).

false negatives) and shift their cognitive focus from labor-intensive manual extraction to better assessing the quality and relevance of the information examined (reducing the false positives). As a case in point, within the allotted four-hour time period, TheoryOn users were able to find, on average, 35.2 synonymous constructs, 35.2 antecedents, and 18 consequents, as well as integrate 13.8 theories—all nearly double compared with EBSCOhost and Google Scholar users. This finding is consistent with our prediction that IT artifacts that disembed behavioral knowledge from large-scale publications can allow users to focus on more value-added activities. The results of our subsequent qualitative applicability check further reinforced and underscored the speed, efficiency, and connection value proposition of TheoryOn during the processing stage of information-seeking behavior.

We also conducted an error analysis of TheoryOn users to understand the system bottleneck. On the one hand, failing to extract relevant constructs/relationships could result in false negatives in users' search results. In time-sensitive situations, examining all the results retrieved by TheoryOn may be challenging for users. This could explain why the user search recalls in Table 3 were lower than the method extraction recalls in Table 1. On the other hand, erroneous constructs/relationships undoubtedly led to some false positives in the search results. However, the users were able to assess and filter out many false positives via manual correction and refinement, which might explain why the users' search precision results were higher than the method extraction precision results presented earlier. Nonetheless, overall, the results of the user experiments suggest that TheoryOn has demonstrated its capabilities of lessening the cognitive load of manually processing knowledge and reducing false positives and false negatives in the scholarly information-seeking process. A future extension for this user experiment is allowing participants to combine and switch between whatever tools they may

choose, such as Google Scholar, EBSCO, and TheoryOn, to yield valuable insights into the particular steps in the information-seeking process in which TheoryOn is most valuable.

### Perceived Utility

According to Table 4, across four tasks related to our proposed system functionalities, we found no significant difference in task experience (TE1–TE4;  $p > 0.05$ ), but the perceived usefulness of TheoryOn for finding synonymous constructs, antecedents, and consequents and for extending theories was significantly better than that of EBSCOhost and Google Scholar, with a difference of 0.72 to 1.69 points on a seven-point Likert scale. Regarding overall utility perception at the system level, TheoryOn was considered to be significantly easier to use (EU) and useful (PU), whereas the behavioral intention to use the system (BI) was marginally significant. This marginal significance is likely due to TheoryOn not being publicly accessible at the time of the experiment; therefore, it was difficult for users to predict whether or not they would access the system in the next six months, which is the time frame used in the BI items.

### Applicability Check

We also conducted an applicability check to evaluate our system's *importance*, *accessibility*, and *suitability* to practitioners (Lukyanenko et al. 2019; Rosemann and Vessey 2008). We recruited 10 academic researchers at the assistant to full professor levels through an announcement to an academic listserv. The advertised inclusion criteria specified that they had to be social or behavioral researchers; had to hold a position equivalent to U.S. titles of assistant, associate, or full professor;

**Table 4. Perceived Usefulness Comparison of TheoryOn, EBSCOhost, and Google Scholar**

<b>Construct</b>	<b>TheoryOn (n = 18)</b>		<b>EBSCOhost (n = 17)</b>			<b>Google Scholar (n = 17)</b>		
	<b>Mean</b>	<b>SD</b>	<b>Mean</b>	<b>SD</b>	<b>Diff (t-stat)</b>	<b>Mean</b>	<b>SD</b>	<b>Diff (t_stat)</b>
PU	5.92	0.73	5.01	1.01	3.04**	4.25	1.37	4.52***
EU	6.21	0.58	5.47	1.28	2.21*	5.78	1.27	2.14*
BI	5.57	1.21	4.84	1.26	1.74	6.57	2.24	-1.64
PU1	6.11	0.54	5.21	0.94	3.54**	5.37	1.42	2.07*
PU2	5.90	0.75	5.44	1.03	2.14*	4.82	1.46	2.77**
PU3	6.44	0.60	4.85	1.44	4.30***	4.96	1.45	4.01***
PU4	5.67	0.99	4.72	1.54	2.17**	4.89	1.57	2.42*
TE1	5.00	1.19	4.61	1.30	0.93	5.55	1.10	-1.41
TE2	5.69	0.92	5.24	1.14	1.29	5.06	1.29	1.66
TE3	5.26	1.35	4.82	1.24	0.99	4.61	1.53	1.34
TE4	4.22	1.46	4.63	1.47	-0.82	5.04	1.44	-1.67

**Notes:** 1. \*p < 0.050; \*\*p < 0.010; \*\*\*p < 0.001.

2. PU: Perceived usefulness of the system; EU: ease of use of the system; BI: behavioral intention to use the system. PU1–4 are the perceived usefulness for each task. TE1–4 are the prior experience with each of the tasks; diff (t-stat) is the t statistics of EBSCOhost or Google Scholar compared with TheoryOn.

had to have published at least five academic articles, and had to be available for two 1.5-hour time slots.

The participants were engaged in two surveys, two one-hour NGT sessions, and a one-hour session and hands-on information search tasks for exposure to TheoryOn. The applicability check revealed 14 steps in the scholarly information-seeking process. For each step, the participants were asked to identify supporting IT artifacts. After being exposed to TheoryOn and completing the information retrieval tasks, the participants were asked to re-examine the information-seeking process and identify steps in which TheoryOn could be a significant help. The detailed process and materials are shown in Appendix E. The NGT sessions were recorded, transcribed, and coded, and the results are summarized in Table 5. In general, the applicability check shed light on how scholarly information-seeking is conducted through three phases (searching, accessing, and processing), highlighted the potential value of construct-oriented search (and TheoryOn) during the processing phase, and touched on the potential for systems, such as TheoryOn, to complement existing options in the search phase.

Specifically, TheoryOn was considered *important* and *useful* for the scholarly information-seeking, especially in the processing phase. TheoryOn saves considerable processing time by readily visualizing theoretical models. Regarding accessibility, the participants applauded the user-friendly and intuitive interface: “wonderful to have a tool to visually support ontology construction” and “very interesting and useful—

especially the graphic visualization.” Regarding suitability, the participants felt that TheoryOn could be especially useful and suitable for novice information seekers, especially those getting into a new field. Moreover, some participants felt that TheoryOn could help experienced researchers validate their understanding of a familiar field, refresh themselves on recent developments, and improve the overall quality of their scholarly pursuits. Furthermore, some participants noted that TheoryOn can support the peer-review process and ensure review quality.

Additionally, they also commented on its complementarity to existing academic support IT artifacts. For example, they pointed out that “Google Scholar gave us coverage, but TheoryOn gave us precision,” and “TheoryOn has the potential to be implemented within the university library system.” Collectively, the applicability check validated the three phases of the information-seeking process, identified the stage in which TheoryOn could be especially helpful, and illustrated its importance, accessibility, and suitability.

## Discussion

In the following, we discuss the design science contribution of our paper by highlighting the accomplishments of the BOLT framework, TheoryOn instantiation, multifaceted evaluation, as well as the generalizability of our proposed design artifacts. Finally, we discuss the potential impact of using the proposed

**Table 5. Summary of the Applicability Check for TheoryOn**

Information-Seeking Behaviors	Nominal Group Technique-Derived Information-Seeking Process	Supporting IT Artifacts	Quotes Related to TheoryOn
Searching	<ul style="list-style-type: none"> <li>• Formulate the problem/phenomenon</li> <li>• Identify the research questions</li> <li>• Identify the search terms</li> <li>• Search relevant articles</li> <li>• Screen for inclusion</li> <li>• Search articles related to the seed articles</li> </ul>	Google Scholar Web of Science Medline Journal and association portals (AIS) ABI-Inform	<p>“By identifying which papers are similar or redundant, TheoryOn can find new publications that I previously neglected.”</p> <p>“TheoryOn could give doctoral students a decent start. It provides a quick and holistic view of a new area.”</p> <p>“It could be a validation tool for reviewers to see whether a meta-analysis or literature review paper did a good job covering all the relevant papers.”</p> <p>“It can work with citation management software, such as Mendeley, to accomplish a comprehensive solution to manage all related papers in a field.”</p> <p>“If you start with a new research question, it is a very good tool to explore and synthesize the relevant research.”</p>
Accessing	<ul style="list-style-type: none"> <li>• Access information systems or library portals</li> </ul>	Web browser	
Processing	<ul style="list-style-type: none"> <li>• Search the relevant keywords from selected articles</li> <li>• Annotate relevant arguments in articles</li> <li>• Discover contexts, variables, and theories</li> <li>• Extract citations</li> <li>• Synthesize arguments, variables, relations, theories, data, and findings</li> <li>• Categorize articles by usefulness and relevance</li> <li>• Build the discourse of the arguments and hypotheses</li> </ul>		<p>“TheoryOn really speeds up everything! It automatically extracts hypotheses, constructs, relationships, and models. So it facilitates synthesizing findings very well.”</p> <p>“The most significant impact that TheoryOn has is six words: speeding up the evaluation of relevant papers. Traditional systems just present the abstracts. But you know, judging the relevance of a paper is more than its abstract. We need to look into variables, models, and findings, which TheoryOn has conveniently provided to us.”</p> <p>“The system tremendously saves us time! This is very important. This morning, I was sitting in a panel and heard someone talking about conducting a literature review of six hundred papers. The most challenging part is to codify these papers. Because TheoryOn extracts all relevant component parts, researchers can concentrate on [improving] review quality rather than manually codifying papers.”</p> <p>“When I look at those models extracted by TheoryOn, I might start to think, hmm ... these relationships are missing. That triggers me to identify new research gaps.”</p> <p>“TheoryOn can help highlight the key variables and constructs from the paper. It can also help me identify the most influential authors and papers—especially when I start a new domain.”</p> <p>“TheoryOn’s ability to pull all the papers and models together and extract all the relevant pieces is amazing!”</p> <p>“TheoryOn can help me link the constructs and save a lot of time. It just automatically does it!”</p> <p>“TheoryOn can help me build my own model. It can creatively suggest new papers or new models because it could find similar constructs between different papers.”</p> <p>“If you already know the field, it helps you refine the research question, validate your understanding, and prioritize the most important papers.”</p>

design artifacts to mitigate the knowledge inaccessibility problem in behavioral research.

**BOLT Framework.** Following Walls et al. (1992), we proposed a BOLT design framework to offer concrete prescriptions for building artifacts capable of extracting specific ontology components related to behavioral knowledge disembedding. The method evaluation results demonstrated the superiority of the state-of-the-art prescriptions offered by the meta-design to support the nuances and complexities associated with the meta-requirements of BOLT. Furthermore, these results collectively underscored the feasibility of adopting the concept-centric perspective (Weber 2012) to disembed behavioral knowledge advocated by BOLT, where the extraction of hypotheses and constructs are critical precursors.

**TheoryOn System.** The BOLT-guided TheoryOn system and its underlying extraction methods constitute important proof-of-concept artifacts. TheoryOn handily outperformed existing ontology learning systems and search engines. In particular, the randomized user experiment results showed that participants using TheoryOn attained F-measures that were 37% to 121% higher for all tasks, relative to the EBSCOhost and Google Scholar full-text search engines. Our applicability check shed light on the scholarly information-seeking process about when, to whom, and how construct-centric search engines might be beneficial, as well as the value proposition of tools such as TheoryOn. Overall, these results highlight the ability of BOLT-guided instantiation—TheoryOn—to extract behavioral knowledge from texts and to enhance information-seeking outcomes for behavioral researchers, verifying the importance of employing a multifaceted evaluation solution to demonstrate the practical value of TheoryOn.

**Multifaceted Evaluation.** Consistent with design principles (Hevner et al. 2004), we used a multifaceted evaluation to rigorously test each component of the proposed IT artifacts. The data mining experiments, randomized user experiment, and qualitative applicability check collectively offer additional empirical and qualitative insights that contribute to the academic literature on knowledge inaccessibility and information seeking in two ways:

- (1) *Intelligent Text Analytics Can Alleviate Knowledge Inaccessibility.* Our randomized user experiment showed that TheoryOn enabled its users to attain significantly better precision and recall, and access behavioral knowledge in an accurate and comprehensive manner. Prior work on the knowledge inaccessibility problem has largely focused on the comprehensiveness/recall problem, and our study confirmed the extent of this problem

(Larsen and Bong 2016)—EBSCOhost and Google Scholar users were only able to retrieve between 9.8% and 34.7% percent of constructs on a fairly small article testbed (i.e., one favorable to higher recall rates). Interestingly, the user study also revealed lower precision rates. On three of the four tasks, the EBSCOhost and Google Scholar users were 6% to 49% lower on precision. This finding suggests that the bandwidth freed up by TheoryOn's automated assistance allows users to shift their cognitive focus from labor-intensive manual extraction to information quality and relevance examination, hence reducing false positives. Future design research on the knowledge inaccessibility problem should consider both precision and recall metrics as important considerations for artifact construction.

- (2) *Empirical Evidence that BOLT Systems Are Possible, Practical, and Valuable for Enhancing Information-Seeking.* The randomized user experiment and applicability check empirically revealed how the phases proposed by the information seeking literature (Meho and Tibbo 2003) are facilitated by the BOLT systems. Specifically, our randomized user experiment demonstrated that automatic behavioral knowledge extraction allows users to search for more articles (searching phase) and process more information in an accurate manner (processing phase). In addition, our qualitative applicability check validated the phases of the information seeking process and highlighted the potential value of implementing BOLT systems with existing artifacts to enhance the searching and processing phases. As far as we know, this article represents the first extensive examination of behavioral information-seeking processes and the potential for new, enabling design artifacts.

**Generalizability.** Our design artifacts could be applied to multiple behavioral and social disciplines such as behavioral medicine, psychology, education, and economics. They are also generalizable to NLP research (Abbasi and Chen 2008; Abbasi et al. 2019; Lau et al. 2012) as well as problem contexts and design solutions at the intersection of data, theory, and ML (Maass et al. 2018) in three ways:

- (1) *Importance of Taking a Concept-Centric Perspective.* The BOLT framework espouses the concept-centric perspective (Weber 2012), which showed that by focusing on effectively extracting hypotheses and constructs, the complex task of disembedding behavioral knowledge becomes viable. This simple and powerful idea of identifying key position statements and concepts nested within those statements can be generalized to many additional contexts such as philosophy and law,

allowing for the development of robust IT artifacts for retrieving “locked” information and knowledge.

(2) *Deep Learning Methods for Complex NLP.* The NLP research in IS has been dominated by topic categorization and sentiment polarity classification (Abbasi and Chen 2008; Abbasi et al. 2018; Lau et al. 2012; Zimbra et al. 2018). From an NLP perspective, these are relatively straightforward binary or multi-class classification problems (although accuracies for sentiment polarity detection remain challenging in certain domains). With the dramatic growth of a variety of user-generated text sources, methods capable of tackling more complex NLP problems such as knowledge extraction from behavioral data are at a premium (Ahmad et al. 2019). The results of our deep learning methods, fused with domain-specific features in a hybridized architecture, shed light on tackling complex NLP tasks in other fields such as biomedical text mining.

(3) *Holistic Evaluation for Design at the Intersection of Data, Theory, and Machine Learning.* Evaluating design artifacts at the intersection of data, theory, and ML is particularly tricky (Maass et al. 2018; Prat et al. 2015). Our work is an example of such artifacts: the BOLT framework and TheoryOn instantiation rely on multiple behavioral and ontology learning theories, involve complex ML algorithms, and address structured and unstructured data throughout the design process. The empirical findings of our multifaceted evaluation solution revealed that a combination of data mining experiments, randomized user experiment, and qualitative applicability checks could help researchers reconcile competing approaches, identify design bottlenecks, and evaluate design solutions from diverse perspectives in this particular design context.

**Impact of Mitigating Knowledge Inaccessibility.** With the aid of our BOLT-guided TheoryOn search engine and combined with conventional search engines such as Google Scholar or EBSCOhost, the scholarly information-seeking process could be better supported, and the knowledge inaccessibility problem in behavioral research could be significantly mitigated. Specifically, with better awareness of existing constructs and relationships (as illustrated by high recalls in the user experiment), researchers are less likely to reinvent constructs or relationships already introduced by others, reducing wasted and redundant efforts as well as marginal research. Consequently, it would be easier to build a cumulative research tradition to ensure the persistent development and progression of a research discipline. Furthermore, by reducing manual effort on information processing,

researchers could improve the agility of their research topics and quickly respond to environmental changes. This research agility and efficiency could lead to profound monetary and societal benefits (e.g., speeding up behavioral intervention design for depression).

## Conclusions and Future Directions ■■■

Our contributions are threefold. First, we propose a BOLT design framework to guide the development of systems capable of behavioral knowledge disembedding and knowledge inaccessibility alleviation. Second, we instantiate our framework into a search engine artifact, TheoryOn, to show the applicability of the framework. TheoryOn also incorporates deep learning methods coupled with a composite kernel SVM to effectively extract hypotheses and constructs and their relations. Finally, through a series of data mining experiments, a randomized user experiment, and a qualitative applicability check, we offer additional empirical and qualitative insights that contribute to the academic literature on NLP research, design at the intersection of data, theory, and ML, information-seeking behaviors, and knowledge inaccessibility.

The level of success with which the hypothesis extraction, variable extraction, and relationship extraction were shown to work, and the improvements to which it led in a search experiment and applicability check, bodes well for the future. The solid performance of our design artifacts shows that future work is likely to be able to perform at such levels that behavioral knowledge disembedding will become the only option imaginable for evaluating past evidence. In fact, over the past 12 months, purely through word of mouth, the system has already garnered an impressive amount of usage. We believe these usage statistics would be further enhanced after a professional upgrade of the UI and UX interface (Kumar et al. 2004).

- **Engagement:** Over 4,000 engaged users who performed an average of 11 major actions per session, with an average session duration of nearly 5½ minutes, and who in total ran over 17,500 unique construct searches.
- **Reach:** These engaged users came from 459 academic institutions across 125 countries, with over 75% of users coming from Europe and Asia.

In this era of profound digital transformation, automation is disrupting various manual processes. Our proposed BOLT framework could have the potential to enable much more

accurate literature search, automatic literature review, and automatic meta-analysis, as well as enable us to chart future directions for these disciplines more efficiently. We expect to work with experts in the biological and computer sciences to further refine and improve the framework proposed here and believe that the IS discipline is the natural home for this kind of work because of our understanding of design science, behavioral approaches, and NLP.

## Acknowledgments

We thank the U.S. National Science Foundation for partial research support under the following grants: SBE-0965338, IIS-1816504, CCF-1629450, BDS-1636933, IIS-1553109, and IIS-1236970. We thank the University of Colorado; the work was also partially supported by the Center for Business Analytics at the University of Virginia. We also thank Grace Xiong, Bill Su, and students affiliated with the Human Behavior Project at the University of Colorado for excellent research assistance.

## References

Abbasi, A., and Chen, H. 2008. "CyberGate: A Design Framework and System for Text Analysis of Computer-Mediated Communication," *MIS Quarterly* (32:4), pp. 811-837.

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

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

Ahmad, F., Abbasi, A., Li, J., Dobolyi, D., Netemeyer, R., Clifford, G., and Chen, H. 2020. "A Deep Learning Architecture for Psychometric Natural Language Processing," *ACM Transactions on Information Systems* (38:1), Article 6.

Ajzen, I. 1991. "The Theory of Planned Behavior," *Organizational Behavior and Human Decision Processes* (50:2), pp. 179-211.

Arazy, O., Kumar, N., and Shapira, B. 2010. "A Theory-Driven Design Framework for Social Recommender Systems," *Journal of the Association for Information Systems* (11:9), pp. 455-490.

Arnulf, J. K., Larsen, K. R., Martinsen, Ø. L., and Bong, C. H. 2014. "Predicting Survey Responses: How and Why Semantics Shape Survey Statistics on Organizational Behaviour," *PloS One* (9:9), p. e106361.

Arnulf, J. K., Larsen, K. R., Martinsen, Ø. L., and Egeland, T. 2018. "The Failing Measurement of Attitudes: How Semantic Determinants of Individual Survey Responses Come to Replace Measures of Attitude Strength," *Behavior Research Methods* (50), pp. 2345-2365.

Ask, K., and Granhag, P.A. 2005. "Motivational Sources of Confirmation Bias in Criminal Investigations: The Need for Cognitive Closure," *Journal of Investigative Psychology and Offender Profiling* (2:1), pp. 43-63.

Asim, M. N., Wasim, M., Khan, M. U. G., Mahmood, W., and Abbasi, H. M. 2018. "A Survey of Ontology Learning Techniques and Applications," *Database: The Journal of Biological Databases and Curation* (2018:1).

Baron, R. M., and Kenny, D. A. 1986. "The Moderator-Mediator Variable Distinction in Social Psychological Research: Conceptual, Strategic, and Statistical Considerations," *Journal of Personality and Social Psychology* (51:6), pp. 1173-1182.

Beel, J., and Gipp, B. 2010. "On the Robustness of Google Scholar against Spam," in *Proceedings of the 21<sup>st</sup> ACM Conference on Hypertext and Hypermedia*, New York: ACM Press, pp. 297-298.

Berger, A. L., Pietra, V. J. D., and Della, P. S. A. 1996. "A Maximum Entropy Approach to Natural Language Processing," *Computational Linguistics* (22:1), pp. 39-71.

Biemann, C. 2005. "Ontology Learning from Text: A Survey of Methods," *LDV-Forum 2005* (20:2), pp. 75-93.

Bodenreider, O. 2004. "The Unified Medical Language System (UMLS): Integrating Biomedical Terminology," *Nucleic Acids Research* (32:Suppl. 1), pp. D267-D270.

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), Article 131.

Buitelaar, P., Cimiano, P., and Magnini, B. (eds.). 2005. *Ontology Learning from Text: Methods, Evaluation and Applications*, Amsterdam: IOS Press.

Bunge, M. A. 1977. "Emergence and the Mind," *Neuroscience* (2:4), pp. 501-509.

Bunge, M. A. 1979. *Treatise on Basic Philosophy: Ontology II: A World of Systems*, Dordrecht, Holland: D. Reidel Publishing Company.

Burges, C. J. 1998. "A Tutorial on Support Vector Machines for Pattern Recognition," *Data Mining and Knowledge Discovery* (2:2), pp. 121-167.

Bushman, B. J., and Wells, G. L. 2001. "Narrative Impressions of Literature: The Availability Bias and the Corrective Properties of Meta-Analytic Approaches," *Personality and Social Psychology Bulletin* (27:9), pp. 1123-1130.

Cho, K., Van Merriënboer, B., Bahdanau, D., and Bengio, Y. 2014. "On the Properties of Neural Machine Translation: Encoder-Decoder Approaches," *arXiv:1409.1259*.

Choudhury, V., and Karahanna, E. 2008. "The Relative Advantage of Electronic Channels: A Multidimensional View," *MIS Quarterly* (32:1), pp. 179-200.

Collins, M., and Duffy, N. 2002. "New Ranking Algorithms for Parsing and Tagging: Kernels over Discrete Structures, and the Voted Perceptron," in *Proceedings of the 40<sup>th</sup> Annual Meeting of the Association for Computational Linguistics*, Stroudsburg, PA: Association for Computational Linguistics, pp. 263-270.

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.

Corley, K. G., and Gioia, D. A. 2011. "Building Theory about Theory Building: What Constitutes a Theoretical Contribution?," *Academy of Management Review* (36:1), pp. 12-32.

Cortes, C., and Vapnik, V. 1995. "Support-Vector Networks," *Machine Learning* (20:3), pp. 273-297.

Cristianini, N., and Shawe-Taylor, J. 2000. *An Introduction to Support Vector Machines and Other Kernel-Based Learning Methods*, New York: Cambridge University Press.

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

Drymonas, E., Zervanou, K., and Petrakis, E. G. 2010. "Unsupervised Ontology Acquisition from Plain Texts: The Ontogain System," in *Proceedings of the Natural Language Processing and Information Systems, and 15<sup>th</sup> International Conference on Applications of Natural Language to Information Systems*, Berlin: Springer-Verlag, pp. 277-287.

Ellis, D. 1989. "A Behavioural Approach to Information Retrieval System Design," *Journal of Documentation* (45:3), pp. 171-212.

Eyre, T. A., Ducluzeau, F., Sneddon, T. P., Povey, S., Bruford, E. A., and Lush, M. J. 2006. "The Hugo Gene Nomenclature Database, 2006 Updates," *Nucleic Acids Research* (34:Suppl. 1), pp. D319-D321.

Faure, D., and Poibeau, T. 2000. "First Experiments of Using Semantic Knowledge Learned by ASIUM for Information Extraction Task Using Intex," in *Proceedings of the First International Conference on Ontology Learning ECAI-2000 Workshop*, Aachen, Germany: CEUR-WS.org, pp. 7-12.

Fellbaum, C. 1998. "A Semantic Network of English Verbs," in *WordNet: An Electronic Lexical Database*, C. Fellbaum (ed.), Cambridge, MA: MIT Press, pp. 153-178.

Friedman, N., Geiger, D., and Goldszmidt, M. 1997. "Bayesian Network Classifiers," *Machine Learning* (29:2-3), pp. 131-163.

Gefen, D., Endicott, J. E., Fresneda, J. E., Miller, J. L., and Larsen, K. R. 2017. "A Guide to Text Analysis with Latent Semantic Analysis in R with Annotated Code: Studying Online Reviews and the Stack Exchange Community," *Communications of the Association for Information Systems* (41:1), pp. 450-496.

Gill, T. G., and Hevner, A. R. 2013. "A Fitness-Utility Model for Design Science Research," *ACM Transactions on Management Information Systems* (4:2), pp. 5:1-5:24.

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.

Hobbs, J., and Riloff, E. 2010. "Information Extraction," in *Handbook of Natural Language Processing*, N. Indurkha and F. J. Damerau (eds.), Boca Raton, FL: CRC Press, pp. 511-532.

Hochreiter, S., and Schmidhuber, J. 1997. "Long Short-Term Memory," *Neural Computation* (9:8), pp. 1735-1780.

Huang, Z., Xu, W., and Yu, K. 2015. "Bidirectional LSTM-CRF Models for Sequence Tagging," *arXiv:1508.01991*.

Im, G., and Straub, D. 2012. "Building Cumulative Tradition in Organization Science: A Methodology for Utilizing External Validity for Theoretical Generalization," Unpublished Paper, Georgia State University.

Jiang, X., and Tan, A. H. 2010. "CRCTOL: A Semantic Based Domain Ontology Learning System," *Journal of the Association for Information Science and Technology* (61:1), pp. 150-168.

Keen, P. 1980. "MIS Research: Reference Disciplines and a Cumulative Tradition," in *Proceedings of the 10<sup>th</sup> International Conference on Information Systems*, Copenhagen, Denmark, October 18-20, 2006, pp. 9-18.

Kim, Y. 2014. "Convolutional Neural Networks for Sentence Classification," *arXiv:1408.5882*.

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

Krosgaard, M. A., Brodt, S. E., and Whitener, E. M. 2002. "Trust in the Face of Conflict: The Role of Managerial Trustworthy Behavior and Organizational Context," *Journal of Applied Psychology* (87:2), pp. 312-319.

Kumar, R. L., Smith, M. A., and Bannerjee, S. 2004. "User Interface Features Influencing Overall Ease of Use and Personalization," *Information & Management* (41:3), pp. 289-302.

Lafferty, J., McCallum, A., and Pereira, F. 2001. "Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data," in *Proceedings of 18<sup>th</sup> International Conference on Machine Learning*, C. E. Brodley and A. P. Danyluk (eds.), San Francisco: Morgan Kaufmann Publishers, pp. 282-289.

Landis, J. R., and Koch, G. G. 1977. "The Measurement of Observer Agreement for Categorical Data," *Biometrics* (33:1), pp. 159-174.

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.

Larsen, K. R., Hekler, E. B., Paul, M. J., and Gibson, B. S. 2020. "Improving Usability of Social and Behavioral Sciences' Evidence: A Call to Action for a National Infrastructure Project for Mining Our Knowledge," *Communications of the Association for Information Systems* (46).

Larsen, K. R., Hovorka, D. S., West, J. D., and Dennis, A. R. 2019. "Understanding the Elephant: A Discourse Approach to Corpus Identification for Theory Review Articles," *Journal of the Association for Information Systems* (20:7), pp. 887-927.

Larsen, K. R., Michie, S., Hekler, E. B., Gibson, B., Spruijt-Metz, Aher, D., Cole-Lewis, H., Bartlett Ellis, R. J., Hesse, B., Moser, R. P., and Yi, J. 2017. "Behavioral Change Interventions: The Potential of Ontologies for Advancing Science and Practice," *Journal of Behavioral Medicine* (40:1), pp. 6-22.

Larsen, K. R., Voronovich, Z. A., Cook, P. F., and Pedro, L. W. 2013. "Addicted to Constructs: Science in Reverse?," *Addiction* (108:9), pp. 1532-1533.

Lau, R. Y., Liao, S. S., Wong, K.-F., and Chiu, D. K. 2012. "Web 2.0 Environmental Scanning and Adaptive Decision Support for Business Mergers and Acquisitions," *MIS Quarterly* (36:4), pp. 1239-1268.

LeCun, Y., Bottou, L., Bengio, Y., and Haffner, P. 1998. "Gradient-Based Learning Applied to Document Recognition," *Proceedings of the IEEE* (86:11), pp. 2278-2324.

Lindén, K., and Puitulainen, J. 2004. "Discovering Synonyms and Other Related Words," *Proceedings of CompuTerm 2004: 3<sup>rd</sup> International Workshop on Computational Terminology*, Geneva, Switzerland: COLING, pp. 63-70.

Luo, Y., Uzuner, Ö., and Szolovits, P. 2016. "Bridging Semantics and Syntax with Graph Algorithms—State-of-the-Art of Extracting Biomedical Relations," *Briefings in Bioinformatics* (18:1), pp. 160-178.

Lukyanenko, R., Evermann, J., and Parsons, J. 2014. "Instantiation Validity in IS Design Research," in *Proceedings of the International Conference on Design Science Research in Information Systems*, M. C. Tremblay, D. VanderMeer, M. Rothenberger, A. Gupta, and V. Yoon (eds.), Heidelberg: Springer, pp. 321-328.

Lukyanenko, R., Parsons, J., Wiersma, Y., and Maddah, M. 2019. "Expecting the Unexpected: Effects of Data Collection Design Choices on the Quality of Crowdsourced User-Generated Content," *MIS Quarterly* (43:2), pp. 623-647.

Ma, X., and Hovy, E. 2016. "End-to-End Sequence Labeling Via Bi-Directional LSTM-CNNs-CRF," *arXiv:1603.01354*.

Maass, W., Parsons, J., Purao, S., Storey, V. C., and Woo, C. 2018. "Data-Driven Meets Theory-Driven Research in the Era of Big Data: Opportunities and Challenges for Information Systems Research," *Journal of the Association for Information Systems* (19:12), pp. 1253-1273.

Maedche, A., and Staab, S. 2000. "Mining Ontologies from Text," in *Proceedings of the International Conference on Knowledge Engineering and Knowledge Management*, Berlin: Springer, pp. 189-202.

March, S. T., and Smith, G. 1995. "Design and Natural Science Research on Information Technology," *Decision Support Systems* (15:4), pp. 251-266.

Maynard, D., Funk, A., and Peters, W. 2009. "Using Lexico-Syntactic Ontology Design Patterns for Ontology Creation and Population," in *Proceedings of the 2009 International Conference on Ontology Patterns* (Volume 516), Aachen, Germany: CEUR-WS.org, pp. 39-52.

McCandless, M., Hatcher, E., and Gospodnetic, O. 2010. *Lucene in Action: Covers Apache Lucene 3.0*, Stamford, CT: Manning Publications Co.

McMillan, J. J., and White, R. A. 1993. "Auditors' Belief Revisions and Evidence Search: The Effect of Hypothesis Frame, Confirmation Bias, and Professional Skepticism," *Accounting Review* (68:3), pp. 443-465.

Meho, L. I., and Tibbo, H. R. 2003. "Modeling the Information Seeking Behavior of Social Scientists: Ellis's Study Revisited," *Journal of the American Society for Information Science and Technology* (54:6), pp. 570-587.

Mikolov, T., Karafiat, M., Burget, L., Černocký, J., and Khudanpur, S. 2010. "Recurrent Neural Network Based Language Model," in *Proceedings of the 11<sup>th</sup> Annual Conference of the International Speech Communication Association*, Baixas, France: International Speech Communications Association, pp. 1045-1048.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., and Dean, J. 2013. "Distributed Representations of Words and Phrases and Their Compositionality," in *Proceedings of the 26<sup>th</sup> International Conference on Neural Information Processing Systems*, C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Q. Weinberger (eds.), Lake Tahoe, NV, pp. 3111-3119.

Miller, G. A. 1995. "Wordnet: A Lexical Database for English," *Communications of the ACM* (38:11), pp. 39-41.

Morita, T., Fukuta, N., Izumi, N., and Yamaguchi, T. 2006. "DODDLE-OWL: A Domain Ontology Construction Tool with Owl," in *Proceedings of the First Asian Conference on the Semantic Web*, Berlin: Springer-Verlag, pp. 537-551.

Muller, K.-R., Mika, S., Ratsch, G., and Scholkopf, B. 2001. "An Introduction to Kernel-Based Learning Algorithms," *IEEE Transactions on Neural Networks* (12:2), pp. 181-201.

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.

Nickerson, R. S. 1998. "Confirmation Bias: A Ubiquitous Phenomenon in Many Guises," *Review of General Psychology* (2:2), pp. 175-220.

Oliveira, A., Pereira, F. C., and Cardoso, A. 2001. "Automatic Reading and Learning from Text," in *Proceedings of the International Symposium on Artificial Intelligence*, Kolhapur, India.

Park, J., Cho, W., and Rho, S. 2007. "Evaluation Framework for Automatic Ontology Extraction Tools: An Experiment," *OTM Confederated International Conferences "On the Move to Meaningful Internet Systems"*, Berlin: Springer, pp. 511-521.

Parsons, J., and Wand, Y. 2013. "Extending Principles of Classification from Information Modeling to Other Disciplines," *Journal of the Association for Information Systems* (14:4), pp. 245-273.

Peffers, K. 2002. "Perishable Research and the Need for a New Kind of IS Journal," *Journal of Information Technology Theory and Application* (4:1), p. v.

Pennington, J., Socher, R., and Manning, C. 2014. "Glove: Global Vectors for Word Representation," *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing*, Stroudsburg, PA: Association for Computational Linguistics, pp. 1532-1543.

Popper, K. 1959. *The Logic of Scientific Discovery*, New York: Basic Books.

Prat, N., Comyn-Wattiau, I., and Akoka, J. 2015. "A Taxonomy of Evaluation Methods for Information Systems Artifacts," *Journal of Management Information Systems* (32:3), pp. 229-267.

Quan, T. T., Hui, S. C., Fong, A. C. M., and Cao, T. H. 2004. "Automatic Generation of Ontology for Scholarly Semantic Web," *International Semantic Web Conference*, Berlin: Springer-Verlag, pp. 726-740.

Quirchmayer, G., Basl, J., You, I., Xu, L., and Weippl, E. 2012. *Multidisciplinary Research and Practice for Information Systems: International Cross Domain Conference and Workshop on Availability, Reliability, and Security*, Prague, Czech Republic.

Rabiner, L. 1989. "A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition," in *Proceedings of the IEEE* (77:2), pp. 257-286.

Rai, A. 2017. "Editor's Comments: Avoiding Type III Errors: Formulating IS Research Problems That Matter," *MIS Quarterly* (41:2), pp. iii-vii.

Rosemann, M., and Vessey, I. 2008. "Toward Improving the Relevance of Information Systems Research to Practice: The Role of Applicability Checks," *MIS Quarterly* (32:1), pp. 1-22.

Salton, G. 1989. *Automatic Text Processing: The Transformation, Analysis, and Retrieval of Information by Computer*, Boston: Addison-Wesley Longman Publishing Co., Inc.

Schryen, G., Benlian, A., Rowe, F., Shirley, G., Larsen, K., Petter, S., Paré, G., Wagner, G., Haag, S., and Yasasin, E. 2017. "Literature Reviews in IS Research: What Can Be Learnt from the Past and Other Fields?," *Communications of the Association for Information Systems* (41:1), Article 30.

Sharman, R., Kishore, R., and Ramesh, R. (eds.). 2007. *Ontologies: A Handbook of Principles, Concepts and Applications in Information Systems*, New York: Springer Science & Business Media.

Sombatsrisomboon, R., Matsuo, Y., and Ishizuka, M. 2003. "Acquisition of Hypernyms and Hyponyms from the WWW," in *Proceedings of the 2<sup>nd</sup> International Workshop on Active Mining*, Japan.

Soper, D. S., and Turel, O. 2015. "Identifying Theories Used in North American IS Research: A Bottom-Up Computational Approach," in *Proceedings of the 48<sup>th</sup> Hawaii International Conference on System Sciences*, Los Alamitos, CA: IEEE Computer Society Press, pp. 4948-4958.

Spell, C. S. 2001. "Management Fashions—Where Do They Come From, and Are They Old Wine in New Bottles?," *Journal of Management Inquiry* (10:4), pp. 348-373.

Strube, M., Rapp, S., and Müller, C. 2002. "The Influence of Minimum Edit Distance on Reference Resolution," in *Proceedings of the ACL-02 Conference on Empirical Methods in Natural Language Processing* (Volume 10), Stroudsburg, PA: Association for Computational Linguistics, pp. 312-319.

Szafranski, M., Grandvalet, Y., and Rakotomamonjy, A. 2010. "Composite Kernel Learning," *Machine Learning* (79:1-2), pp. 73-103.

Tan, S. S., Lim, T. Y., Soon, L.-K., and Tang, E. K. 2016. "Learning to Extract Domain-Specific Relations from Complex Sentences," *Expert Systems with Applications* (60), pp. 107-117.

Tang, D., Qin, B., Feng, X., and Liu, T. 2015. "Effective LSTMS for Target-Dependent Sentiment Classification," *arXiv:1512.0110*.

Trinh, T. P., Molla, A., and Peszynski, K. 2012. "Enterprise Systems and Organizational Agility: A Review of the Literature and Conceptual Framework," *Communications of the Association for Information Systems* (31:1), pp. 167-193.

Vargas-Vera, M., Domingue, J., Kalfoglou, Y., Motta, E., and Buckingham Shum, S. 2001. "Template-Driven Information Extraction for Populating Ontologies," in *Proceedings of the IJCAI'01 Workshop on Ontology Learning*, Seattle, WA.

Vapnik, V. N. 1998. *Statistical Learning Theory*, New York: John Wiley & Sons.

Venkatesh, V., and Morris, M. G. 2000. "Why Don't Men Ever Stop to Ask for Directions? Gender, Social Influence, and Their Role in Technology Acceptance and Usage Behavior," *MIS Quarterly* (24:1), pp. 115-139.

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.

Walls, J. G., Widmeyer, G. R., and El Sawy, O. A. 1992. "Building an Information System Design Theory for Vigilant EIS," *Information Systems Research* (3:1), pp. 36-59.

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

Webster, J., and Watson, R. T. 2002. "Analyzing the Past to Prepare for the Future: Writing a Literature Review," *MIS Quarterly* (26:2), pp. xiii-xxiii.

White, R. 2013. "Beliefs and Biases in Web Search," in *Proceedings of the 36<sup>th</sup> International ACM SIGIR Conference on Research and Development in Information Retrieval*, New York: ACM Press, pp. 3-12.

Wong, W., Liu, W., and Bennamoun, M. 2012. "Ontology Learning from Text: A Look Back and Into the Future," *ACM Computing Surveys* (44:4), Article 20.

Yadav, V., and Bethard, S. 2018. "A Survey on Recent Advances in Named Entity Recognition from Deep Learning Models," in *Proceedings of the 27<sup>th</sup> International Conference on Computational Linguistics*, Stroudsburg, PA: Association for Computational Linguistics, pp. 2145-2158.

Yousafzai, S. Y., Foxall, G. R., and Pallister, J. G. 2007. "Technology Acceptance: A Meta-Analysis of the TAM: Part 1," *Journal of Modelling in Management* (2:3), pp. 251-280.

Zhou, G., Su, J., Zhang, J., and Zhang, M. 2005. "Exploring Various Knowledge in Relation Extraction," in *Proceedings of the 43<sup>rd</sup> Annual Meeting on Association for Computational Linguistics*, Stroudsburg, PA: Association for Computational Linguistics, pp. 427-434.

Zhou, G., Qian, L., and Fan, J. 2010. "Tree Kernel-Based Semantic Relation Extraction with Rich Syntactic and Semantic Information," *Information Sciences* (180:8), pp. 1313-1325.

Zhou, P., Qi, Z., Zheng, S., Xu, J., Bao, H., and Xu, B. 2016. "Text Classification Improved by Integrating Bidirectional LSTM with Two-Dimensional Max Pooling," *arXiv:1611.06639*.

Zimbra, D., Abbasi, A., Zeng, D., and Chen, H. 2018. "The State-of-the-Art in Twitter Sentiment Analysis: A Review and Benchmark Evaluation," *ACM Transactions on Management Information Systems* (9:2), Article 5.

## About the Authors

**Jingjing Li** is an assistant professor of Information Technology in the McIntire School of Commerce at the University of Virginia. She received her Ph.D. in Information Systems from the Leeds School of Business, University of Colorado. Her research interests relate to artificial intelligence and big data analytics, with applications in search engine, recommender system, healthcare, behavioral ontology learning, consumer behavior, and public policy making. Her

research has been published in elite journals in the Information Systems, Marketing, and Management fields, and earned her the INFORMS Design Science Award, WITS Best Paper Award, and WITS Best Prototype Award. She has received grants from the U.S. National Science Foundation, Amazon Web Services, and Microsoft. When she worked as a scientist at Microsoft, she proposed and implemented a variety of machine learning solutions to tackle complex business and societal problems.

**Kai R. Larsen** is an associate professor of Information Systems in the division of Organizational Leadership and Information Analytics, Leeds School of Business, University of Colorado Boulder. He is a courtesy faculty member in the Department of Information Science of the College of Media, Communication and Information, a Research Advisor to Gallup, and a Fellow of the Institute of Behavioral Science. Kai is most known for providing a practical solution to Edward Thorndike's (1904) Jingle Fallacy and for his contributions to the semantic theory of survey response (STSR), which holds that results of surveys using attitude scales primarily measure the linguistic relationships between survey questions.

**Ahmed Abbasi** is the Giovannini Endowed Chaired Professor in the Department of IT, Analytics, and Operations in the Mendoza College of Business at the University of Notre Dame. He received his Ph.D. in Information Systems from the Artificial Intelligence Lab at the University of Arizona, and an M.B.A. and B.S. degrees in Information Technology from Virginia Tech. Ahmed has 20 years of experience pertaining to AI and predictive analytics, with applications in health, text mining, online fraud and security, and social media. His research has been funded by over a dozen grants from the U.S. National Science Foundation and industry partners such as Amazon Web Services, eBay, Microsoft, and Oracle. He has also received the IEEE Technical Achievement Award, INFORMS Design Science Award, and IBM Faculty Award for his work at the intersection of machine learning and design. Ahmed has published over 100 articles in top journals and conferences, and won the AIS top publication and *MIS Quarterly* best paper awards. His work has been featured in various media outlets, including the *Wall Street Journal*, *Harvard Business Review*, the Associated Press, WIRED, and CBS. Ahmed serves on the editorial board for various IS, ACM, and IEEE journals.

# Appendix A

## Alternative Techniques for BOLT Framework

**Table A1. Additional Techniques for BOLT Framework**

Outputs	Task	Techniques	Description
Hypotheses (Terms)	Hypothesis Extraction	Maximum Entropy (ME)	ME (Berger et al. 1996) directly estimates a conditional probability of class labels given input features. It treats hypothesis extraction as a sentence classification problem. Y reflects whether a sentence is a hypothesis and X contains the input features that describe a sentence.
		Naïve Bayes (NB)	NB (Friedman et al. 1997) is a generative classifier, which tries to learn an optimal joint probability of input features and class label. Similar to ME, NB treats hypothesis extraction as a sentence classification problem. However, its performance is subjective to the ratio between positive and negative cases.
Constructs	Variable Extraction	Conditional Random Fields (CRF)	CRF (Lafferty et al. 2001) is a discriminative sequence labeler that directly estimates conditional probability. It takes a complex set of linguistics features to predict labels that are dependent on each other. For variable extraction, variables are tagged according to IOB schema. CRF then tries to find the best IOB sequence to identify a variable in a sentence.
		Hidden Markov Model (HMM)	HMM (Rabiner 1989) is a generative sequence labeler that directly estimates the joint probability. It is subjective to the influence of the class labels, and usually needs more assumptions to make the estimation tractable. Similar to CRF, it tries to find the best IOB sequence to extract variables.
Theoretical Relationships	Theoretical Relationship Extraction	Semantic Template	Semantic Template (Vargas-Vera et al. 2001) utilizes lexical and syntactical features to detect ontological relations through extraction rules.
		Syntactic Structure Analysis	Syntactic structure analysis and dependency analysis (Sombatsrisomboon et al. 2003) examines syntactic and dependency information to discover terms and their relations at the sentence level.
Construct Hierarchy	Synonymous Relation Identification	Clustering	Clustering (Lindén and Pittulainen 2004) employs measures of similarity to assign terms into groups. The clusters could be organized as a hierarchy.

## Appendix B

### Rule-Based Hypothesis Extraction Rules

The hypothesis formatting rules are identified as follows. A training data set was used to create an initial set of extraction rules. Next, the rules were iteratively refined by examining the results on a validation set. The refinement process concluded when a reasonable F-measure, precision, and recall were attained on the validation set (F-measure = 92.98%; precision = 96.94%; recall = 89.34%). Consequently, five extraction rules were identified and represented as regular expressions:

- (1) Hypothesis starts with “H” and a number (e.g. H1) or an alphabet (e.g. H1a)  
‘^H[0-9]{1,2}[a-zA-Z]?[: \.]? \*[A-Z].+\\$’
- (2) Hypothesis starts with “Hypothesis” and a number (e.g. Hypothesis 1) or an alphabet (e.g. Hypothesis 1a)  
‘^Hh[Yy][Pp][Oo][Tt][Hh][Ee][Ss][Ii][Ss] ?[0-9]{0,2}[a-zA-Z]?[: \.]? \*[A-Z].+\\$’
- (3) Hypothesis starts with “Proposition” and a number (e.g. Proposition 1) or an alphabet (e.g. Proposition 1a)  
‘^Pp[Rr][Oo][Pp][Oo][Ss][Ii][Tt][Ii][Oo][Nn] ?[0-9]{0,2}[a-zA-Z]?[: \.]? \*[A-Z].+\\$’
- (4) Hypothesis starts with “Hypothesis” and a number followed by “H”+ a number (Hypothesis 1 (H1))  
‘^Hh[Yy][Pp][Oo][Tt][Hh][Ee][Ss][Ii][Ss] ?[0-9]{0,2}[a-zA-Z]? ?H[0-9]{0,2}[a-zA-Z]? \*[[: \.]]? ?[A-Z].+\\$’
- (5) Hypothesis starts with “Hypothesis” and a number followed by “H”+ a number wrapped by parentheses (Hypothesis 1 (H1))  
‘^Hh[Yy][Pp][Oo][Tt][Hh][Ee][Ss][Ii][Ss] ?[0-9]{0,2}[a-zA-Z]? ?(H ?[0-9]{0,2}[a-zA-Z]? ?)[: \.]? \*[A-Z].+\\$’

# Appendix C

## Composite Kernel Function in Relation Extraction

We used SVM with a composite kernel function to extract the derived binary theoretical relationships from hypotheses (Kitchens et al. 2018). Formally, a training data set  $(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$  consists of variable pairs, where  $x_i$  is a feature vector describing a particular variable pair (e.g., the number of words contained between variable instances), and  $y_i$  is a binary label with 1 indicating “having that particular relation” (e.g., main effect), using a one-against-all scheme. We need to find optimal hyperplanes when

$$\begin{aligned} \text{Maximize: } & \text{margin} = \frac{2}{\|w\|} \\ \text{Subject to: } & y_i(w \cdot x_i + b) - 1 \geq 0 \end{aligned} \quad (1)$$

The Lagrange Function Formulation is used to solve this minimization problem, and we get the dual problem

$$\begin{aligned} \text{Maximize: } & W(\alpha) = \sum_{i=1}^n \alpha_i - \frac{1}{2} \sum_{i=1}^n \sum_{j=1}^n \alpha_i \alpha_j y_i y_j K(x_i \cdot x_j) \\ \text{subject to: } & \alpha_i \geq 0, i = 1, 2, \dots, n \\ & \sum_{i=1}^n \alpha_i y_i = 0 \end{aligned} \quad (2)$$

where  $\alpha_i$  is the dual variable, and  $K(x_i \cdot x_j)$  is the kernel function of the feature vectors of two variables to measure the similarity between two feature vectors by mapping them to a higher-dimensional space, and can be tailored to incorporate domain-specific knowledge (Burges 1998; Muller et al. 2001). Specifically, composite kernels are well suited to incorporate broad, relevant features while reducing the risk of over-fitting (Collins and Duffy 2002; Szafranski et al. 2010). An effective composite kernel is commonly represented as a linear combination of several types of kernels (Zhou et al. 2010). For our first kernel, a linear feature-based kernel, we adopted a comprehensive feature list from Zhou et al. (2005) to build flat feature vectors representing the linguistic patterns between two variables. The details are in Table C1.

$$FK(s_i, s_j) = \frac{\langle s_i, s_j \rangle}{\sqrt{\langle s_i, s_i \rangle \langle s_j, s_j \rangle}} \quad (3)$$

The second kernel utilized the augmented subtree  $ST$  generated in the first step and computed the parse tree similarity as the number of common substructures. Specifically, for each pair of variables in a hypothesis, the kernel function  $TK(st_i, st_j)$  measures the similarity between  $ST_i$  and  $ST_j$ , computed by comparing all their tree substructures, where a substructure is defined as any subgraph containing more than one node (Collins and Duffy 2002). Formally, let  $I_k(st_i)$  denote the presence of the  $k^{\text{th}}$  tree substructure in  $ST_i$  (where  $I_k(st_i) = 1$  if the  $k^{\text{th}}$  in  $st_i$ ). Accordingly,  $ST_i$  can be represented as a binary vector  $I(x_i) = (I_1(x_i), \dots, I_n(x_i))$  representing the presence of different tree substructures. Hence,  $TK(st_i, st_j)$  can be computed as two times the number of common substructures in  $ST_i$  and  $ST_j$ , divided by the total number of substructures in  $ST_i$  and  $ST_j$ .

$$TK(st_i, st_j) = \frac{2 \sum_{k=1}^n (I_k(st_i) I_k(st_j))}{\sum_{k=1}^n (I_k(st_i) + I_k(st_j))} \quad (4)$$

Finally, the composite kernel (CK) function is created to fully exploit the diverse linguistic patterns manifested in structural and linear feature-based cues, taking the following form:

$$CK = TK + \tau FK \quad (5)$$

In this equation,  $\tau$  is the parameter to adjust the relative weight assigned to the feature vector kernel and tree kernel functions, and it is determined through validation.

**Table C1. Features Used in Feature Vector-Based Relation Extraction (Zhou et al. 2005)**

Category	Attribute	Feature	Description
Words	Words of both mentions	WM1, WM2	Bag-of-words in M1, Bag-of-words in M2
	Words between the two mentions	WBNULL	No words in between
		WBFL	The word between M1 and M2 when there is only one word in between.
		WBF, WBL	The first (WBF) and last (WBL) word between M1 and M2 when at least two words in between
		WBO	Words except for first/last words between M1 and M2
	Words before M1	BM1F	First word before M1
		BM2L	Second word before M1
	Words after M2	AM1F	First word after M2
		AM2L	Second word after M2
Counts	Mentions between pair	#MB1	Number of construct mentions in between
	Mentions before pair	#BM1	Number of construct mentions before this pair
	Mentions after pair	#AM1	Number of construct mentions after this pair
	Words between pair	#WB	Number of words in between
	Words before pair	#WBF	Number of words before M1
	Words after pair	#WAF	Number of words after M2
Phrases	Phrases between the pair	CPHBNULL	No phrase in between
		CPHBFL	The phrase head when only one phrase in between
		CPHBF	First phrase head when at least two phrases in between
		CPHBL	Last phrase head when at least two phrases in between
		CPHBO	Phrase heads except for first and last phrase in between
	Phrases before M1	CPHBM1F	First phrase head before M1
		CPHBM1L	Second phrase head before M1
	Phrases after M2	CPHAM2F	First phrase head after M2
		CPHAM2L	Second phrase head after M2
Parse Tree	Features from tree	PTP	Path of phrase labels connecting M1 and M2 in tree
Order	Occurrence order of mentions	M1<M2	M1 precedes M2
		M1>M2	M2 precedes M1

# Appendix D

## Detailed Description of Randomized User Experiment

To evaluate TheoryOn's ability to retrieve large-scale behavioral knowledge, we selected two full-text search engines, Google Scholar and the Business Source Complete database powered by EBSCOhost. Both of them represented, at the time of the experiment, the longest uninterrupted period of full-text coverage for *MIS Quarterly*, *Information Systems Research*, and *Journal of Applied Psychology* (1990–2009). Users in all three groups were guided to search within the same period and journals.

A total of 52 information systems and organizational behavior Ph.D. students worldwide were randomly assigned to one of the three experimental groups (TheoryOn, EBSCOhost, or Google Scholar). To ensure randomization, we conducted ANOVA tests on the three groups based on demographics such as *age*, *years of work experience*, and *years in a Ph.D. program*, and found that none of them were significantly different ( $p > 0.05$ ). However, we found one indicator, *prior experience with search engine*, was significantly higher for the EBSCOhost and Google Scholar users than for the TheoryOn group ( $p < 0.05$ ). This difference suggests an advantage for the traditional full-text system: *ceteris paribus*, the Google Scholar and EBSCOhost groups would be likely to perform better than the TheoryOn group due to the former's greater system familiarity.

### Tasks

To test TheoryOn system's utility, we designed four tasks for each participant to complete: *synonymous construct search*, *construct pair search*, *antecedents and consequents search*, and *theory integration*, each of which is a common scholarly information task for behavioral research. All four tasks were related to one theory, the technology acceptance model (TAM), in order to demonstrate a natural progression of knowledge acquisition, curation, and integration in an information-seeking process. TAM was selected due to high awareness, which again set up a context in which users of Google Scholar and EBSCOhost were given every opportunity to perform at their peak. The detailed task description is in Table D1.

For each task, the participants were given an example of a construct, a construct pair, or a theory, along with necessary details such as construct definition and sample items. In order to familiarize the participants with the functionalities of TheoryOn, EBSCOhost, and Google Scholar, a short video tutorial (3–5 minutes) was given for each task. The participants were required to complete each task in less than an hour. On average, participants self-reported that the synonymous construct search, construct pair search, antecedents and consequents search, and theory integration tasks took 42.33, 23.93, 40.01, and 46.01 minutes, respectively.

### Evaluation Methods

Multiple evaluation metrics can provide a comprehensive view of the utility and fitness of a design artifact (Hevner et al. 2004). Therefore, we evaluated TheoryOn's performance using the two metrics of objective and perceptual evaluations, where the objective evaluation compared the construct, article, and theory retrieval performance including precision and recall (Salton 1989), and the subjective evaluation examined the perceived utility of the artifact.

### Objective Metrics

Each participant's submission was compared against a carefully constructed gold standard set using precision, recall, and  $F_1$ -measure. Precision was then calculated as the number of correctly identified constructs or articles divided by the total number of constructs or articles retrieved by each participant. Recall was calculated as the number of correctly identified constructs or articles divided by the total number in the gold standard set. The  $F_1$ -measure was the harmonic mean of precision and recall. Specifically, recall can be considered to be a metric to measure confirmation bias (e.g., Ask and Granhag 2005; McMillan and White 1993).

The gold standard for each task was rigorously constructed by a team of two experienced faculty researchers, three doctoral students, and four senior research assistants (research assistants had at least 500 hours of experience in construct extraction from behavioral articles). Starting with the constructs described in Table D1, all the relevant constructs and their residing articles from the three focal journals—*MIS Quarterly*, *Information Systems Research*, and *Journal of Applied Psychology* from 1990 to 2009—were identified. The inclusion decision was judged

Table D1. Tasks in the Randomized Experiment.		
Task Description/Submission	Construct/Definition	Sample of Items
<p>Synonymous Construct Search: Find as many synonymous constructs as possible for <b>Perceived Usefulness</b>.</p> <p><b>Submission:</b> Synonymous constructs along with their article information.</p>	<p>Perceived Usefulness (Davis 1989; Venkatesh et al. 2003): The degree to which a person believes that using a particular system would enhance his or her job performance.</p> <p><b>N:</b> 123 constructs for perceived usefulness.</p>	<ul style="list-style-type: none"> <li>Using the system in my job would enable me to accomplish tasks more quickly.</li> <li>Using the system would improve my job performance.</li> <li>Using the system in my job would increase my productivity.</li> <li>Using the system would enhance my effectiveness on the job.</li> <li>I would find the system useful in my job.</li> </ul>
<p>Construct Pair Search: Find as many articles as possible that contain both <b>Perceived Usefulness</b> (see Task 1 Definition) and <b>Trust</b>, including articles that contain both of their synonymous counterparts.</p> <p><b>Submission:</b> Articles containing both constructs (including synonymous constructs).</p>	<p>Trust (Choudhury and Karahanna 2008): A user's beliefs about the reliability, credibility, and accuracy of information gathered through the web.</p> <p><b>N:</b> 10 articles containing perceived usefulness and trust.</p>	<ul style="list-style-type: none"> <li>I would have greater confidence in the explanations provided by such web sites than in those offered by an agent.</li> <li>I would trust the validity of quotes provided by this web site more than those provided by an agent.</li> <li>I believe such a web site would provide more objective recommendations than an agent would provide.</li> <li>I would feel more confident purchasing the policy through the web than through an agent.</li> </ul>
<p>Antecedents and Consequents Search: For the construct <b>Perceived Usefulness</b>, find as many immediate antecedents and consequents as possible (i.e., the constructs that are hypothesized to directly influence or be influenced by <b>Perceived Usefulness</b>).</p> <p><b>Submission:</b> Immediate antecedents and consequents with their article information.</p>	<p>See Task 1</p> <p><b>N:</b> 95 immediate antecedents and 55 consequents.</p>	See Task 1
<p>Theory Integration: Extend the original Technology Acceptance Model (TAM) (Davis 1989) by integrating relevant hypothetical relationships through constructs synonymous with <b>Perceived Usefulness</b>, <b>Perceived Ease of Use</b>, and <b>Behavioral Intention to Use</b>. Each article must contain <b>Behavioral Intention</b> and at least one construct from <b>Perceived Usefulness</b> and <b>Perceived Ease of Use</b>.</p> <p><b>Submission:</b> Articles that integrated with TAM and an expanded TAM model diagram.</p>	<p>Perceived Ease of Use (Davis 1989; Venkatesh et al. 2003): The degree to which a person believes that using a system would be free of effort.</p> <p><b>N:</b> 39 articles containing either Perceived Usefulness or Ease of Use.</p> <p>Behavioral Intention to Use (Davis 1989; Venkatesh et al. 2003): Participant's intention to use the technology.</p>	<ul style="list-style-type: none"> <li>Learning to operate the system would be easy for me.</li> <li>I would find it easy to get the system to do what I want it to do.</li> <li>My interaction with the system would be clear and understandable.</li> <li>I would find the system to be flexible to interact with.</li> <li>I would find the system easy to use.</li> </ul> <ul style="list-style-type: none"> <li>I intend to use the system in the next n months.</li> <li>I predict I would use the system in the next n months.</li> <li>I plan to use the system in the next n months.</li> </ul>

by two independent research teams, and the final adjudication was determined by the team with experienced faculty researchers. The second column in Table D1 states the number of constructs/articles in the gold standard for each of the four tasks.

### Perceptual Metrics

Following the evaluation guidelines by Hevner et al. (2004) and Gill and Hevner (2013), we adapted multiple scales to evaluate the perceptual utility of TheoryOn. Specifically, immediately after completing each task, the participants were asked to report the helpfulness of the system on a four-item *Usefulness* scale adapted from Venkatesh et al. (2003). In addition, for each task, we asked three questions related to *Task*

Experience to make sure there were no significant differences in task familiarity between the two experimental groups. After the participants completed all tasks, they were asked to report on their perception of three TAM constructs adapted from Davis (1989) and Venkatesh et al. (2003): a four-item *Perceived Usefulness* scale, a four-item *Perceived Ease of Use* scale, and a three-item *Behavioral Intention to Use* scale. All of the scales were operationalized using a seven-point Likert scale.

## Appendix E

### Detailed Process and Results for Applicability Check

To evaluate the relevance of TheoryOn, we applied Rosemann and Vessey's (2008) applicability check approach with additional guidance from Lukyanenko et al. (2019) to develop understanding around the needs of our researcher-as-practitioner community. Per Rosemann and Vessey's instructions, the applicability check was conducted as a part of the research cycle and TheoryOn was left unchanged after the check.

The applicability check was conducted to evaluate our system's "*importance, accessibility, and suitability* to practitioners" (Rosemann and Vessey 2008, pp. 9-10). We recruited 10 academic researchers at the assistant- to full-professor levels through an announcement to an academic listserv.<sup>9</sup> Advertised inclusion criteria specified that they had to be social or behavioral researchers; had to hold a position equivalent to U.S. titles of assistant, associate, or full professor; had to have published at least five academic papers; and be available at for two 1.5-hour time slots. Each participant was rewarded with a \$100 debit card for their time. No performance conditions beyond participation in all three hours of the process were specified.

The participants were engaged in five different elements:

1. A pre-applicability check survey
2. Applicability Check Step 1: a one-hour nominal group technique (NGT) session where the participants were engaged to share their information seeking process.
3. Applicability Check Step 2: a one-hour process whereby the participants were first introduced to the design artifact (TheoryOn) and then worked on their own to understand it and to explore how it could potentially help them in their information seeking process.
4. Applicability Check Step 3: an online survey about their beliefs regarding the design artifact after first exposure.
5. Applicability Check Step 4: a one-hour NGT session where the participants were asked to first individually reflect on their experiences with TheoryOn and then prompted to think through how it could be used in their own information seeking process.

#### Pre-applicability Check Survey

A pre-survey revealed the average participant to have 17.4 years of academic experience, having published 27 journal articles and 39 conference proceedings. One was an assistant professor, four were associate professors, and five were full professors. When responding on a Likert scale, all but one participant agreed or strongly agreed with statements that they felt comfortable doing literature reviews related to behavioral constructs, understood behavioral constructs, and were comfortable in their use. The last participant disagreed with all three statements. This level of familiarity and comfort with behavioral constructs may reflect the Information Systems (IS) discipline's focus on such constructs. A pre-applicability check survey then asked each participant to (1) list their information seeking steps, (2) explain what information systems or library portals they used for each of the information-seeking steps, and (3) list which steps of the information-seeking process could not be helped by existing information systems. The responses to each question were summarized and shared with the participants in summarized form:

- (1) Major Information Seeking steps:
  - (a) Starting: keywords, variables/constructs, phenomenon/topic, theory, paper
  - (b) Expansion: references or causal relationship (main, moderation, mediation or control variables)
  - (c) Extraction: manually extract and read through papers or studies
- (2) Information systems or library portals:
  - (a) Google Scholar, Google
  - (b) EBSCO host, ABI/Inform, university libraries, Proquest, Medline, Web of Science, AIS, journal portals
  - (c) Endnote/Mendeley/Excel

<sup>9</sup>An eleventh researcher had signed up for but withdrew on the day of the applicability check.

(3) Steps of the process conducted with no IS support:

- (a) Formulation of the research question
- (b) Identification of relevant theories and frameworks as well as core constructs
- (c) Search literature based on relevant theories
- (d) Screen for inclusion
- (e) Extract data from papers
- (f) Synthesize findings:
  - (i) Arguments
  - (ii) Causal relations
  - (iii) Hypotheses
  - (iv) Antecedent variables
  - (v) Mediating variables
  - (vi) Dependent variables
- (g) After the research results are available, verify with the core reviewed articles
- (h) Revise the discourse of arguments and update the review

The major discovery from the survey and a discussion with the participants was how few IS tools beyond full-text search and reference managers were used by the participants. Eight major steps in the research process were mentioned by one or multiple researchers as being fully conducted without technology support.

### ***Applicability Check Step 1: Understanding the Information Seeking Process***

While our original plan called for using the pre-survey to split participants into groups based on epistemological differences, no such differences were found, and the participants were randomly assigned to two groups. The group sessions were recorded to help the researchers understand the context of the written group answers.

The participants were not given any information on the overall goals or artifact design before or during this step. The following are the 14 steps outlined by the two teams:

- Formulate the problem/phenomenon
- Identify research questions
- Identify search terms
- Search relevant articles
- Screen for inclusion
- Search articles related to the seed articles
- Access information systems or library portals
- Search the relevant keywords from selected articles
- Annotate relevant arguments in articles
- Discover contexts, variables, and theories
- Extract citations
- Synthesize arguments, variables, relations, theories, data, and findings
- Categorize articles by usefulness and relevance
- Build discourse of arguments and hypotheses

Once each team had agreed to a set of steps for their information seeking process, they were asked to evaluate each step in terms of the process, with regard to which tools they were currently using.

### ***Applicability Check Step 2: Exposure to Artifact***

Half an hour was set aside for explaining the context and introducing the artifact itself. We started by discussing a few of the numerous IS theories that have received thousands of citations. The problem of construct synonymy (Larsen and Bong 2016) was further discussed. The BOLT framework was briefly discussed before screenshots illustrating the four different types of functionality were outlined along with a screenshot for each: (a) construct search, (b) construct-pair search, (c) theoretically related construct search, and (d) theory integration.

To further familiarize the participants with TheoryOn, a one-page description of TheoryOn's context, objectives, and expected utility was developed. To evaluate the importance, accessibility, and suitability of the design artifact, participants were asked to view a set of four video tutorials and instructed to use the artifact for their own construct review in each of the four areas. The one-page description was followed with instructions for viewing the videos and applying TheoryOn to a problem of they chose (see **Exhibit E1**).

### Exhibit E1. Instructions for Exposure to the IT Artifact

- (a) **Construct Search.** TheoryOn allows users to specify a construct in a search query, only returning articles that contain 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. The figure shows a search for *perceived usefulness* using a combination of keyword and Latent Semantic Analysis search. Retrieved constructs are shown with citation information and the ability to examine definitions, items, and operationalization origins. Users may also begin a new semantic or taxonomic search with the current construct as the starting point. When a theoretical network has been extracted from the paper, it is visualized along with the construct information and the target construct marked in yellow. For more details, watch the video "[TheoryOn: Synonymous Construct Search](#)."
- (b) **Construct-Pair Search.** TheoryOn allows users to specify a construct pair in a search query and only returns articles containing these two constructs. The constructs (marked in yellow) 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](#)."
- (c) **Theoretically Related Construct Search.** This functionality allows inspection of the theoretical models containing a construct of interest (highlighted in yellow) as well as examination of its antecedents and consequents in a list or plot view. TheoryOn takes the first  $n$  papers returned by the construct search and displays the antecedents to the searched-for construct. It then does the same for the consequents. For more details, watch the video "[TheoryOn: Theoretically Related Construct Search](#)."
- (d) **Theory Integration.** All the related theories can be saved in the sorting hierarchy (left panel) 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](#)."

The participants were then assigned an optional “assignment” to complete four information retrieval tasks related finding relevant constructs about the TAM. The tasks include synonymous construct search, construct-pair search, theoretically-related construct search and theory integration. The detailed description of the tasks are in Appendix D.<sup>10</sup> Each participant has one night to complete the tasks. All participants have completed at least one task and two participants have completed all four tasks. After individual exposure to the artifact, the participants were asked to fill out a survey. The survey and the survey results were not shared with the participants; they are described in the section “Applicability Check Step 3: Post-Exposure Survey.”

### Applicability Check Step 3: Post-Exposure Survey

Upon finishing the hands-on exposure to the system videos and the system itself, the respondents were asked to fill out a survey. The survey contained one open-ended question and a common assembly of artifact evaluation constructs: *effort expectancy (ease of use)*, *performance expectancy (usefulness)*, *facilitating conditions*, and *behavioral intention to use*. All Likert-type scales were from Venkatesh et al. (2003).

<sup>10</sup>It would have been ideal to develop a task set different from the randomized user experiment for the applicability check. For instance, “assume that you are revising a paper and try to find sufficient relevant literature from IS and reference disciplines for trust in social media usage...” However, due to the time constraints between sessions associated with the applicability check, a more prolonged, periodic longitudinal field task was not possible. We acknowledge this as a limitation of the study.

## Exhibit E2. Post-Exposure Survey

1. Please tell us your thoughts about this homework and the system you just experienced [open-ended]
2. Effort expectancy:
  - a. My interactions with the system were clear and understandable [7-point Likert]
  - b. It would be easy for me to become skillful at using the system [7-point Likert]
  - c. I would find the system easy to use [7-point Likert]
  - d. Learning to operate the system is easy for me [7-point Likert]
3. Performance expectancy:
  - a. I would find the system useful in my research [7-point Likert]
  - b. Using the system enables me to accomplish tasks more quickly [7-point Likert]
  - c. Using the system increases my productivity [7-point Likert]
  - d. If I use the system, I will increase my chances of getting a raise [7-point Likert]
4. Facilitating conditions:
  - a. I have the resources necessary to use the system [7-point Likert]
  - b. I have the knowledge necessary to use the system [7-point Likert]
  - c. The system is not compatible with other systems I use [7-point Likert]
  - d. A specific person (or group) is available for assistance with system difficulties [7-point Likert]
5. Behavioral intention to use the system:
  - a. I intend to use the system in the next six months [7-point Likert]
  - b. I predict I would use the system in the next six months [7-point Likert]
  - c. I plan to use the system in the next six months [7-point Likert]

Nine participants filled out the survey with high *effort expectancy* scores, suggesting that the system use processes are clear, it was easy to learn how to use, easy to use, and easy to become skillful in its use (mean = 6.22, SD = .71). The *performance expectancy* construct also came in with strong support for the artifact (mean = 5.7, SD = 1.14), but the average for the last question, that the system would increase the participant's chance for a raise (mean = 4.00, SD = 2.12), was much lower and may indicate that a quality literature review process itself is not seen as having much of an effect on salaries. Removal of this question led to strong scores on *performance expectancy* (mean = 6.26, SD = .92).

*Facilitating conditions* showed a split response set in that the first two questions about having the necessary knowledge and resources showed strong support for the system (mean = 6.22, SD = .76). The third question, about whether the system is compatible with other systems in use (mean = 4.22, SD = 1.78, scores reversed), indicates that the Endnote integration may have been seen as helpful by some, but others may have wanted this system better integrated with their favorite search engines. The final question, about having a specific person or group available for assistance with system difficulties (mean = 4.78, SD = 2.05), was higher than expected given that no support system was established for this applicability check. However, this may be reflective of a problem two participants had connecting to the system from their hotel rooms. Two of the authors communicated with the two participants over email, and were able to confirm the problem, after some time, as partly attributable to an overloaded hotel WIFI. Both these participants rated this question as "strongly agree." Finally, *intention to use* the artifact in the next six months (mean = 5.67, SD = 1.43) was somewhat high, but not as high as it could be. Two participants exhibited only middling interest in using the system in the future, pulling the average down from the levels seen for *ease of use* and *usefulness*. One of these two shared during the session the next day that he simply did not do this kind of theory-based construct research anymore, and therefore was unlikely to use the system in the future. The second person who indicated a middling intention to use the system was the same person who, in the pre-applicability check survey, suggested a lack of comfort in doing literature reviews related to behavioral constructs.

Overall, the survey feedback on *effort expectancy* and *performance expectancy* were exceedingly supportive of the system, and on par with or considerably above other artifact tests in design science research.

The qualitative feedback was qualitatively categorized by one author and is reported on below. Four *general comments* were received, suggesting that the participants found the fundamental principles underlying the artifact "great" and "quite interesting." One participant suggested that he thought "this is an amazing software program" and another shared that she thought the artifact was an "excellent system for theory building and literature review. Very creative! Great job!"

Six comments were received related to the *ease of use* of the system. Three of these comments were positive and in line with the *effort expectancy* scores, so they are not discussed. One was negative, suggesting that the artifact was “not very easy to use for me yet.” The last two had specific points to make that may improve the interface:

- “UI a bit awkward for ontology building—maybe keep all the buttons (zoom, scrolling, and add cluster) together?”
- “In ontology building, sometimes highlighting an item caused it to be turned yellow, other times green, other times red. Wasn’t clear what those colors meant.”

Five comments were received about the *performance expectancy* of the artifact. Three were positive but did not add information beyond the high scores on the quantitative part of the survey; however, one of these comments focused on the usefulness of the system for users intending to develop research models or integrate several existing models. One respondent pointed out a specific functionality he liked and also suggested a new feature:

- “I especially like the LSA functionality, which allows finding synonymous constructs; this is especially useful in behavioral sciences like ours. Having said that, it would be great if the system could also allow the conduct of searches based on empirical findings. This could be of significant help for those who conduct theory-testing reviews like meta-analyses and vote-counting reviews.”

Three remaining topics were found in the survey feedback, each with two comments. First, the *visualizations* were lauded: First, the *visualizations* were lauded: “wonderful to have tool to visually support ontology construction” and “very interesting and useful—especially the graphic visualization.” Second, one respondent had two worries about security: “not running HTTPS” and “how is password stored? Can I delete it? Or change it?” Finally, two respondents wanted more journals and data in the final system, as should be the goal in any final implementation of TheoryOn.

### **Applicability Check Step 4: Modified Nominal Group Technique Applicability Check**

The applicability check technique described by Rosemann and Vessey (2008) allows participants to reflect on their individual experiences and beliefs before sharing those with the group to enable shared discussions and group summarization. Exhibit E3 shows the instructions provided the participants, asking them to first work alone then as a group to answer the question of whether TheoryOn might support any of the steps of the information-seeking process.

#### **Exhibit E3. Instructions**

##### **Group 1**

**Name** \_\_\_\_\_

**Instructions:**

Going back to the steps you come up with yesterday, which of the steps do you think TheoryOn might successfully support for you? Are there additional use cases for TheoryOn?

You have 10 minutes to write down your thoughts individually and 15 minutes to discuss within the group. One of the group members should take notes on the discussion and summarize the thoughts. Be prepared to present your group findings to all the participants at the end of the session.

**Note:** Please organize your thoughts in accordance with the step number in the Notes from Session 1

**Notes from Session 1:**

[This section contained a list of the 14 steps found by the two groups in the first session, but each group was only reminded of and responded to their own steps.]

The last half hour of the session was used to address questions related to the interface of TheoryOn before asking the participants to reflect on any compatibility issues and areas of improvement. In response to questions about the user interface, participants had no negative comments, stating that it is “well-designed and well-thought of,” “intuitive,” and “easy to use.”

In response to a question about whether TheoryOn could be used in conjunction with existing information systems such as Google Scholar, they pointed out that “Google Scholar gave us more coverage but TheoryOn gave us more precision,” but that “TheoryOn has a potential to be implemented within the university library system,” and that “Once TheoryOn is seamlessly integrated with some bibliographic software, it could be a powerful tool for us behavioral researchers.” They further suggested that if “TheoryOn is integrated with the subscription services, it will be an overarching tool for us.”

In response to a question about the main areas of improvement for TheoryOn, respondents had the following suggestions:

- “If the system can selectively show the core constructs, that would be great!”
- “Currently, the ranking is not based on citations. It would be great to consider citations.”
- “Because it is a machine learning algorithm, there are some errors. It would be great if the users could edit the results and share them with others.”

While the system actually does use citations to rank search query returns, the other two suggestions are quite reasonable and will be considered for future releases.

We recorded and transcribed all the NGT session. The transcripts are coded by two authors in the research team. The main results are summarized in Table 6. The applicability check shed light on the scholarly information seeking process and how it relates to the three information-seeking phases, highlighted the potential value of construct-oriented search (and TheoryOn) during the *processing* phase, and touched on the potential for systems such as TheoryOn to complement existing options in the *search* phase. After being exposed to TheoryOn, participants in the applicability check demonstrated tremendous excitement and interest. They felt TheoryOn could be especially useful and suitable for novice information seekers, especially those getting into a new field, as it can quickly extract, connect and present relevant theoretical components. Moreover, some participants also felt TheoryOn could help experienced researchers to validate their understanding about a familiar field, refresh on recent developments, and improve the overall quality of their scholarly pursuits. Some participants also noted that the tool could benefit reviewers by helping to maintain quality while adding convenience in the peer-review process. Collectively, the applicability check demonstrates that our instantiation system is important and suitable for scholars in their information seeking process.

Copyright of MIS Quarterly is the property of MIS Quarterly and its content may not be copied or emailed to multiple sites or posted to a listserv without the copyright holder's express written permission. However, users may print, download, or email articles for individual use.