# THEORYON: A DESIGN FRAMEWORK AND SYSTEM FOR UNLOCKING BEHAVIORAL KNOWLEDGE THROUGH ONTOLOGY LEARNING

## ONLINE APPENDICES

## Appendix A. Alternative Techniques for BOLT Framework

**Table A. 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 $P(Y\|X)$ 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 $(Y, X) = P(X\|Y)P(Y)$. 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 2001) is a discriminative sequence labeler that directly estimates conditional probability $P(Y\|X)$. 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 $P(Y, X) = P(X\|Y)P(Y)$. It is subjective to the influence of the class labels $P(Y)$, 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 (Linden and Pittulainen 2004) employs measures of similarity to assign terms into groups. The clusters could be organized as a hierarchy. |

### References

Lafferty, J., McCallum, A., and Pereira, F. 2001. "Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data," in: *Proceedings of 18th International Conference on Machine Learning*. Williamstown, MA: pp. 282-289.

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.

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

Rabiner, L. 1989. "A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition," in: *IEEE Proc.* pp. 257-286.

Vargas-Vera, M., Domingue, J., Kalfoglou, Y., Motta, E., & Buckingham Shum, S. 2001. Template-driven information extraction for populating ontologies. In *IJCAI'01 Workshop on Ontology Learning*, Seattle, WA.

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

Linden, K. and Piitulainen, J. 2004. "Discovering Synonyms and Other Related Words,": *Proc. Intl. Wrkshp Computational Terminology*.

## 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 $n$ 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

$$\text{Maximize:} \quad margin = \frac{2}{||w||} \qquad (1)$$

$$\text{Subject to:} \quad y_i(w \cdot x_i + b) - 1 \geq 0.$$

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

$$\text{Maximize:} \quad W(\alpha) = \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \alpha_j y_i y_j \mathrm{K}(x_i \cdot x_j) \qquad (2)$$

$$\text{subject to:} \quad \alpha_i \geq 0, i = 1, 2, \dots, n$$

$$\sum_{i=1}^{n} \alpha_i y_i = 0,$$

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 $s_i$ representing the linguistic patterns between two variables. The details are in Table C. The kernel function between two feature vectors $s_1$ and $s_2$ was formulated as the dot product

$$FK\ (s_1,\ s_2) = \frac{\langle s_1, s_2 \rangle}{\sqrt{\langle s_1, s_1 \rangle \langle s_2, s_2 \rangle}}\ . \tag{3}$$

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*th tree substructure in $ST_i$ (where $I_k(st_i) = 1$ if the *k*th tree substructure exists in $st_i$). Accordingly, $ST_i$ can be represented as a binary vector $I(x_i) = (I_1(x_i),\ \ldots\ ,I_n(x_i))$ 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 $ST_j$.

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

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:

$$CK = TK + \tau FK \tag{5}$$

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 from the training data.

**Table C. 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 | Number of  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 |

## Reference

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

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 40th Annual Meeting of the Association for Computational Linguistics*. Stroudsburg, PA: Association for Computational Linguistics, pp. 263-270.

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.

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.

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

Zhou, G., Su, J., Zhang, J., and Zhang, M. 2005. "Exploring Various Knowledge in Relation Extraction," in: *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*. 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.

## 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 from programs in the United States and around the globe 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 D.

**Table D. Tasks in the Randomized Experiment.**

| Task Description/Submission | Construct/Definition | Sample of Items |
|---|---|---|
| Synonymous Construct Search:<br>Find as many synonymous constructs as possible for **Perceived Usefulness**<br>**Submission**:<br>Synonymous constructs along with their article information | Perceived Usefulness (Davis 1989; Venkatesh et al. 2003):<br>The degree to which a person believes that using a particular system would enhance his or her job performance.<br>**N:** 123 constructs for perceived usefulness | • Using the system in my job would enable me to accomplish tasks more quickly.<br>• Using the system would improve my job performance.<br>• Using the system in my job would increase my productivity.<br>• Using the system would enhance my effectiveness on the job.<br>• I would find the system useful in my job. |
| Construct Pair Search:<br>Find as many articles as possible that contain both **Perceived Usefulness** (See Task 1 Definition) and **Trust**, including articles that contain both of their synonymous counterparts.<br>**Submission**:<br>Articles containing both constructs (including synonymous constructs) | Trust(Choudhury and Karahanna 2008):<br>A user's beliefs about the reliability, credibility, and accuracy of information gathered through the web.<br>**N:** 10 articles containing perceived usefulness and trust. | • I would have greater confidence in the explanations provided by such web sites than in those offered by an agent.<br>• I would trust the validity of quotes provided by this web site more than those provided by an agent.<br>• I believe such a web site would provide more objective recommendations than an agent would provide.<br>• I would feel more confident purchasing the policy through the web than through an agent. |
| Antecedents and Consequents Search:<br>For the construct **Perceived Usefulness**, find as many immediate antecedents and consequents as possible, i.e., the constructs that are hypothesized to directly influence or be influenced by **Perceived Usefulness**.<br>**Submission**:<br>Immediate antecedents and consequents with their article information | See Task 1<br>**N:** 95 immediate antecedents and 55 consequents. | See Task 1 |
| Theory Integration:<br>Extend the original Technology Acceptance Model (TAM) (Davis 1989) by integrating relevant hypothetical relationships through constructs synonymous with **Perceived Usefulness**, **Perceived Ease of Use**, and **Behavioral Intention to Use**. Each article must contain **Behavioral Intention** and at least one construct from **Perceived Usefulness** and **Perceived Ease of Use**.<br>**Submission**:<br>Articles that integrated with TAM and an expanded TAM model diagram | Perceived Ease of Use (Davis 1989; Venkatesh et al. 2003):<br>The degree to which a person believes that using a system would be free of effort.<br>**N:** 39 articles containing either Perceived Usefulness or Ease of Use | • Learning to operate the system would be easy for me.<br>• I would find it easy to get the system to do what I want it to do.<br>• My interaction with the system would be clear and understandable.<br>• I would find the system to be flexible to interact with.<br>• I would find the system easy to use. |
| | Behavioral Intention to Use (Davis 1989; Venkatesh et al. 2003):<br>Participant's intention to use the technology. | • I intend to use the system in the next n months.<br>• I predict I would use the system in the next n months.<br>• I plan to use the system in the next n months. |

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 D, 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 by two independent research teams, and the final adjudication was determined by the team with experienced faculty researchers. The second column in Table D 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.

## References

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.

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

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

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

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.

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

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.

## Appendix E. Detailed Process and Results for Applicability Check

To evaluate the relevance of TheoryOn, we applied Roseman and Vessey (2008)'s applicability check approach with additional guidance from Lukyanenko et al. (2019) to develop understanding around the needs of our researcher-as-practitioner community. Per Roseman 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" (p. 10). We recruited 10 academic researchers at the assistant- to full-professor levels through an announcement to an academic listserv[1]. 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.

[1] An 11th researcher had signed up for but withdrew on the day of the applicability check.

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.

### E.1. 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 (a) list their information seeking steps, (b) explain what information systems or library portals they used for each of the information-seeking steps, and (c) 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

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.

### E.2. 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.

### E.3. 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 E.1**):

**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."

**Exhibit E.1: Instructions for Exposure to the IT Artifact.**

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[2]. Each participant has one night to complete the tasks. All participants have completed at least one task and two participants have completed all four

[2] 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 discplines 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.

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 Section X.4.

### E.4. 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).

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]

**Exhibit E.2: Post-exposure Survey.**

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 (e.g., Chang 2004)

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.

### E.5. Applicability Check Step 4: Modified Nominal Group Technique Applicability Check

The applicability check technique described by Roseman 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 D.3 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.

**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 minutewes 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.]

**Exhibit E.3. Instructions.**

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 questions 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.

## References

Rosemann, M., and Vessey, I. 2008. “Toward Improving the Relevance of Information Systems Research to Practice: The Role of Applicability Checks,” *MIS Quarterly* (3:1), pp. 1-22.

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*, Forthcoming

Hoehle, H., and Venkatesh, V. 2015. "Mobile Application Usability: Conceptualization and Instrument Development," *MIS Quarterly* (39:2), pp. 435-472.

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.

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