Leveraging Indexical Pragmatics (OFIP) for Search Engine: An Ontology-based Approach

Dapeng Liu
Virginia Commonwealth University
liud22@vcu.edu

Victoria Yoon
Virginia Commonwealth University
vyyoon@vcu.edu

Abstract

The relevance of search results is an important indicator of information retrieval performance. A domain-specific Search Engine (SE), distinct from a general web SE, focuses on a specific segment of online content and may increase search results relevance. Traditional methods to improve domain-specific SE precision heavily depend on query expansion, lexical analysis of texts, and large amounts of training data. These methods suffer from limited effectiveness and efficiency because expanded query terms and coarse language features bring in uncontrollable complexity and increase dimensionality. Our design, leveraging the integrated power of computational syntax, semantics, and indexical pragmatics, proposes an ontology-driven framework that is tailored to work in a dynamic Internet environment without large amounts of manually annotated training data. This article presents our design, that is essential for building a domain-specific SE, and its instantiation in the terrorism domain.
Keywords: Ontology, Indexical pragmatics, Search relevance, Information extraction

1. Introduction

Digital information has been growing at a dramatic pace. In November 2016, 47% of the world's population (about 3.49 billion) are active Internet users who access and simultaneously generate the digital data on the Internet [1]. Text is the dominant data type on the Internet. Much research has been conducted on search engines (SE) that assist users to access the desired information, mining text available on the Internet [2].

While Google is dominating the market of generic SEs, scholars now are paying increasing attention to domain-specific SEs spanning a variety of domains, such as medicinal chemistry [3], scientific publications [4] and job offers [2]. A domain-specific SE, distinct from a general web SE, focuses on a specific segment of online content. Information retrieval with a domain-specific SE has been an enduring and challenging research task - (1) on the one hand, researchers seek to improve recall, incorporating search results from all relevant Web pages on the Internet; (2) on the other hand, they pursue high precision, providing users the relevant search results to satisfy their needs. Domain-specific SEs bear advantages, in comparison to generic SEs, of being aware and exploiting knowledge of the respective domains to improve search performance. Domain-specific SEs are becoming increasingly popular since they provide increased accuracy and extra features that are difficult to obtain with general SEs [5].

Current domain-specific SEs use traditional machine learning methods which heavily depend on query expansion, lexical analysis of texts, and large amounts of training data [2, 6-8]. These methods suffer from limited effectiveness or efficiency because the expanded query term and coarse language features bring in uncontrollable complexity and increase dimensionality. To address these limitations, this study proposes an ontology-driven design that leverages indexical pragmatics to improve the result relevance of domain-specific SEs. Specifically, our solution employs not only basic NLP techniques such as part of speech tagging and parsing for analyzing syntax features, but also semantic web technique to annotate semantic segments which are extracted via advanced NLP tasks such as entity and relation extraction. Further, the technique of coreference resolution, which assists in the deictic fixing of Time, Place and Person, leverages the power of indexical pragmatics to help map expressions that refer to the same entity (e.g., Time, Place, Person, Organization, etc.). In other words, our design leverages the integrated power of computational linguistics – syntax, compositional semantics, and indexical pragmatics for text annotation and interpretation.

HICSS

Ontologies, as a declarative representation of entities referring to the terms in a specific subject area, serve to organize entities that can enhance indexical pragmatics analysis. Our design, which consists of crawling module, feature extraction module, ontology generation module, and classification module, enables effective identification and interoperation of informative text segments. We instantiate our design in the terrorism domain and build the "Terrorism-Domain Search Engine."

The remainder of the paper is organized as follows. Section 2 provides a review of the related works. Section 3 presents the theoretical underpinning for our design choice. Section 4 and 5, drawing upon the Information Systems Design Theory (ISDT), present our research methodology and the architecture of our artifact. Section 6 presents an instantiation of our SE and discusses the major modules. Finally, section 7 discusses the contribution, implications and limitations.

2. Related works

Providing additional value and exploiting knowledge for specific use, domain-specific SEs exist in a variety of fields, including medicinal chemistry [3], song melodies [9], products [10], scientific publications [4], geography space [11, 12], software applications [8], web services [13] , employment [2, 14], transportation [15], and document schema [16]. Hanbury and Lupu [17] define a domain-specific SE as "a search engine that specifies one or more of the following five dimensions: (1) subject areas, (2) modality, (3) users, (4) tasks, (5) tools, techniques and algorithms required to complete the tasks." Scholars and practitioners now have been paying increasing attention to the studies on domain-specific SEs.

The popularly employed mechanisms to improve SE relevance include query expansion [7], lexical analysis of texts [8], classifiers with large amounts of training data [5], and filters which allow the narrowing down of search results [2]. Prior studies also incorporate the technique of semantic web to tackle word disambiguation problems for improving search precision [18]. As some scholars argue for the necessity of domain-specific knowledge to improve SE performance [19], a large number of recently proposed search enhancement tools have utilized the notion of context in domain-specific SEs [20]. Accordingly, domain-specific engines stand a good chance of providing highly relevant results.

Chronologically, domain-specific SEs have been evolving through three main phases of efforts. Early domain-specific SEs mainly employed machine learning methods with a large volume of manually annotated training data to improve SE precision [5]. In the second phase, as NLP techniques have gotten increasing maturity, scholars integrated advanced NLP tasks into SEs to deal with free text, thereby improving SE performance [21]. In the third phase, the technique of semantic web and ontology is incorporated to improve data representation and interpretation, gaining data interoperability [22, 23]. Further, an ontology-driven domain-specific SE is scalable to large knowledge bases and can capture the semantic segments in free text.

While achieving improved SE performance, scholars have continuously worked to develop solutions that require no large amount of manually labeled data and reduce the training cost. For example, Schmidt, Schnitzer and Rensing [2] leverage filters which allow the narrowing down of search results based on pre-defined filter categories. Geng, Yang, Xu and Hua [24] propose a regularization-based algorithm that keeps away laborious labeling tasks and time-consuming training models.

Additionally, the existent solutions of SE have put much effort into leveraging the power of computational semantics and syntax to understand and annotate text information. For example, Widyantoro and Yen [4] developed a PASS (personalized abstract search services) system which is a web-based domain-specific SE for searching abstracts of academic publications. The PASS design focuses mainly on the terms in text document collection, extracting the fuzzy relations between two terms and building the ontology from these term relations [4]. Yang [13] proposes an ontology-supported SE for web services. This ontology-driven design merely leverages the word-level information in the ontology to classify a webpage. More recently, Bouhana, Zidi, Fekih, Chabchoub and Abed [15] propose an ontology-based SE for personalized itinerary search in urban freight transport systems to improve the accuracy of case retrieval and to reduce retrieval time. It calculates textual similarities when processing information at the word and phrase level. However, these studies reveal that existent domain-specific SE solutions that are ontology-driven are not mature enough: they either show no features of domain-specific knowledge when populating ontology or they ignore information regarding indexical pragmatics [25-27].

In this study, we aim to leverage indexical pragmatics to address the aforementioned issues, thereby advancing an ontology-driven domain-specific SE.

3. Theoretical background

In SE for textual information retrieval, the source data is unstructured text in free format. From the perspective of computational linguistics, it is quite difficult for machines to understand the semantic meanings of the unstructured text. The unstructured text needs to be transformed into structured or semi-structured data so that machines can better understand the data and then apply machine learning and similarity calculation algorithms to retrieve relevant textual information. Noticeably, in the process of preparing text data, individual words do not always present explicit semantic meaning that machines can understand; in some domains, such as counterterrorism study, information is sensitive to indexical information.

According to the implicature theory, language meanings go beyond what the texts or speeches say, as the meaning relies on the contextual and conventional meanings of words or sentences [28].
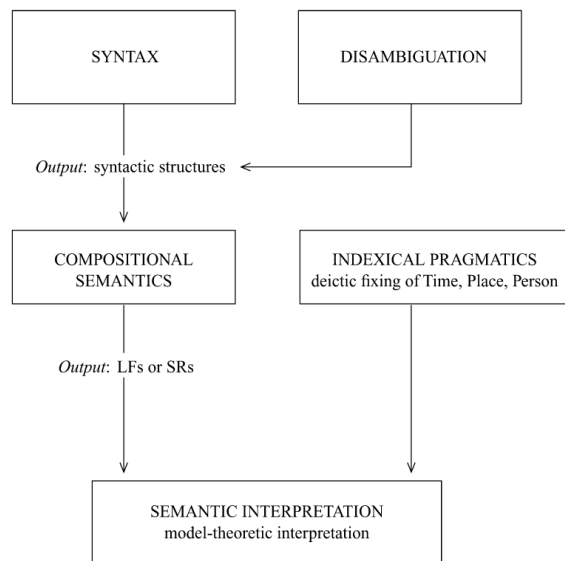

Figure 1. Implicature and Semantics/Pragmatics Interface from [28]

More specifically, computational linguistics revolve around the complex relationship between syntax, compositional semantics, and indexical pragmatics of natural language. The integration of syntax, compositional semantics and indexical pragmatics are emphasized for semantic interpretation [28].

In the case of SE, the solutions should not only employ NLP techniques as part of speech and text tagging and parsing for analyzing syntax features, but also semantic web technique to annotate the parsed

sentences into semantic segments. Additionally, advanced NLP techniques such as entity and relation extraction will help us identify the useful semantic segments. Further, the technique of coreference resolution, emphasizing the power of indexical pragmatics, will be helpful to find and map all expressions that refer to the same entity, such as people, locations, and organizations entities. Therefore, leveraging the integrated power of syntax, semantics, and indexical pragmatics, our design extracts the useful information fragments from the text that can help machines to better understand and interpret the semantic meanings in the text in order to improve information retrieval performance.

4. Research Methodology

The objective of this study is to develop an ontology-driven framework for domain-specific information retrieval, which is designed to improve SE performance (e.g., search relevance). The proposed framework employs the domain knowledge and leverages the integrated power of computational syntax, semantics, and indexical pragmatics for the annotation and interpretation of a text message.

Our study follows the design science methodology [29-31]. More specifically, drawing upon the Information Systems Design Theory (ISDT) [30], our study develops the design artifact and evaluates it regarding the following eight components: (1) purpose and scope, (2) constructs, (3) principles of form and function, (4) artifact mutability, (5) testable propositions, (6) justificatory knowledge, (7) principles of implementation, and (8) expository instantiation. These eight components of ISDT direct the continuous process of capturing, articulating, justifying, and communicating the design knowledge (Table 1).

Table 1. Information systems design theory components

Core components	Description
Purpose and scope	Motivated to address the limitations in existent SE solutions and further improve SE performance, we propose a design artifact, called Ontology-driven Framework leveraging Indexical Pragmatics (OFIP) for domain-specific information retrieval. The proposed design is tailored to work in the dynamic Internet environment.
Constructs	The essential underlying constructs in OFIP include text information

	streaming on the Internet, domain knowledge, informative information granules at various levels, as well as– the linguistic and domain features.
Principle of form and function	Leveraging the integrated power of syntax, semantics, and indexical pragmatics, we adapt the principles and concepts regarding syntax, semantics and pragmatics to the process of text annotation and interpretation. These principles and concepts guide our artifact to extract linguistic and domain features from the input text and transform the text into ontological representation.
Artifact mutability	Our proposed design is scalable. Its utilities can be enhanced by incorporating additional domain knowledge. Further, the artifact is capable of adjusting to the dynamic open Internet within which text information streams in real-time style.
Testable propositions	Drawing on the existent literature and implicature theory, testable propositions are proposed to evaluate the artifact performance.
Justificatory knowledge	We show how OFIP works, by referencing existing literature and the underlying theoretical concepts of computational syntax, semantics, and indexical pragmatics.
Principles of implementation	Guidelines are given on how to adapt the techniques of NLP and semantic web to domain-specific information retrieval.
Expository instantiation	An illustration of working instantiation is provided.

The first component of the ISDT is the purpose and scope of the design. Essentially, design is a process of goal adaption [32]. Accordingly, the design process is goal oriented and should be clearly defined and scoped. The purpose of this study is to propose an ontology-driven framework, leveraging both linguistic features and domain features of text information, for domain-specific information retrieval. The proposed design is tailored to work in the dynamic Internet environment. Our design is fairly grounded in the fact that existent methods heavily employ the coarse natural language features and large amounts of training data to improve the relevance of search results.

Constructs are the elementary units of a design, encompassing the entities of interest [30]. They

should be clearly defined and consistently understood and applied without any ambiguity. The most important underlying constructs in OFIP include the text information streaming on the Internet, domain knowledge, and features of text information at various levels.

The principles of form and function that underpin a design are the proposed organization of its constructs [30]. The principles are often deemed as the "blueprint" of a design, which defines its architecture. In this paper, we adapt the concepts and principles regarding syntax, semantics, and pragmatics to the text information analysis process. These principles organize the workflow and architecture of our artifact, especially guiding our artifact to conduct information extraction and transform a text into the ontological representation which annotates the multi-level linguistic and domain features of the input text.

Artifact mutability is important to design artifacts due to the fact that information systems are always involved in endogenous or exogenous changes [30]. Artifact mutability is an indicator of the generalizability of a design in different situations. Our design is sufficiently general and is applicable with scalability. While our instantiated artifact is built for the terrorism domain information retrieval, it can expand the utilities by incorporating additional domain knowledge. Further, the artifact has the capability of adjusting to the dynamic, open Internet within which text information streams in real-time style.

Testable propositions form the basis of artifact evaluations regarding the artifact's actual efficiency, effectiveness, or utility against its stated ones [30]. An ISDT as a theory should be falsifiable [33]. Drawing on the existent literature and implicature theory, the following testable propositions are proposed to evaluate the artifact performance:

Proposition 1: The precision of OFID will be higher than that of the benchmark framework employing only linguistic features.

Proposition 2: The recall of OFID will be higher than that of the benchmark framework employing only linguistic features.

Proposition 3: The F-measure of OFID will be higher than that of the benchmark framework employing only linguistic features.

Justificatory knowledge relates to the kernel theories employed in a design [30]. As is posited in Kuechler and Vaishnavi [34], kernel theories advise design solutions and demonstrate how the designed artifact is derived from existent justificatory knowledge and should lead to promised outcomes [29]. This study presents how and why our OFIP framework works,

drawing upon existing literature and implicature theory as the underlying kernel theory.

The principles of implementation are the guidelines given to direct the artifact implementation [30]. In this study, the guidelines are given and reflected in the process of our adapting the techniques of NLP and semantic web to the artifact construction.

Expository instantiation refers to the physical instantiation of the proposed design artifact, transforming a theorized design to an instantiated system, which makes artifact evaluation possible [30]. Our design has been instantiated in the context of terrorism domain, with the support of the domain knowledge in Global Terrorism Database (GTD).

5. Ontology-Driven framework leveraging Indexical Pragmatics (OFIP)

Our proposed design embodies an ontology-driven framework for a domain-specific SE, retrieving real-time data relevant to a specific domain as well as extracting and managing real-time knowledge on the Internet. NLP techniques of entity extraction and relation extraction assist in extracting information segments and building up the ontological representation of text knowledge.

Ontology "deals with semantic heterogeneity in structured data" [35]. It is an explicit formal specification of the concepts extracted from a knowledge domain [36]. Ontology achieves increasing popularity in the IS community for knowledge engineering, data interoperation, data integration and so forth in design science research. Additionally, IS literature highlights the effectiveness of ontologies in decision support when ontologies often serve at the center and drive data interoperability and knowledge engineering [37]. In our study, ontology empowers domain-specific information retrieval process to annotate and interoperate concepts in terrorism domain knowledge.

Additionally, our design emphasizes the aspect of the indexical pragmatics of text information. Our proposed framework is designed to overcome the existent limitations inherent to traditional designs, which tend to employ extensive language features that are too coarse to be representative for text classification. Further, our design aims to reduce manual work and decrease complexity and dimensionality of features for classification while raising precision. Figure 2 shows the architecture of our design artifact, consisting of crawling, feature extraction, ontology generation, and text classification modules.
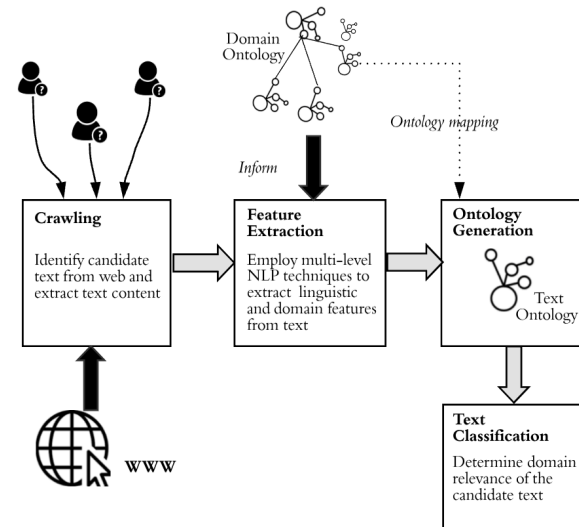
Figure 2. System Architecture

Crawling Module is an automatic crawling agent, informed by a user query, browsing and identifying the candidate text on the Internet. Google search API is a good option for crawling and collecting texts based on keyword query. However, the precision of search results from Google API is criticized by practitioners to be lower than Google website search. In our design, the crawling module runs continuously, identifying, collecting and recording all query-based search results from Google website search; therefore, our artifact will have an equivalent capability of recall, while outperforming in regards to precision. The last stage of crawling is to extract the text content from the web pages identified and retrieved in the prior stages.

Feature Extraction Module, leveraging the integrated power of syntax, semantics, and indexical pragmatics, utilizes a combination of techniques to extract text features. Specifically, our design emphasizes both the indexical information resolved by the technique of coreference resolution [28] and domain-specific features. The feature extraction module goes through a process of coreference resolution, entity extraction, relation extraction, and domain feature extraction.

The technique of coreference resolution, leveraging the power of indexical information, finds and maps all expressions that refer to the same entity in a text (Figure 3). It is an important step for higher level NLP tasks for natural language understanding, especially information extraction [38].

Figure 3. Coreference resolution

Subsequently, the techniques of entity extraction and relation extraction are employed to extract such features as the entities (e.g., people, places) and relations of these entities in the text. For example, the technique of relation extraction can identify and extract the people and their actions (towards other entities) in the text (Figure 4).

Figure 4. E-R recognition and extraction

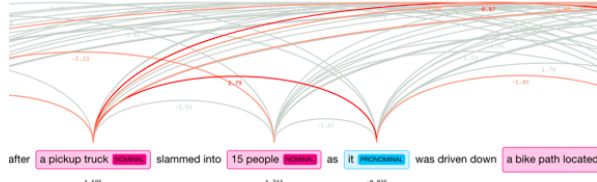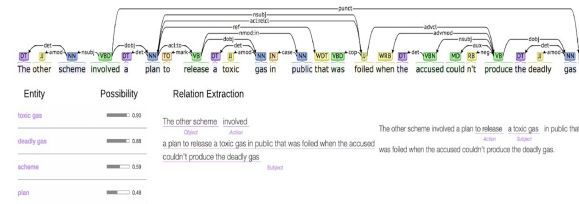Additionally, the domain knowledge, as the reference framework, assists us in extracting domain features from the text and calculating the domain relevance metrics of the retrieved text. For example, a terrorism knowledge base (e.g., Global Terrorism Database), will provide a collection of terrorism domain specific features such as weapon type, attack target type, attack type, facility, terrorism perpetrator, and attack motivation. These domain features can be extracted from a given text and are referenced as domain relevance indicators. Specifically, we transform these textual domain relevance indicators into metrics in the following way. For a given text t, we calculate its numeric domain relevance indicators—its degree of relevance with the domain features. Assuming $f = (c_1, c_2, ... c_j ...)$ represents a feature f consisting of a collection of textual concept c_j, we calculate with the weight function:

$$relevance(t, f) = \sum_{c_j \in f} tf(t, c_j) * weight(c_j, f)$$

,where $tf(t, c_j)$ is the within-text term frequency of a concept c_j in t, $weight(c_j, f)$ denotes the weight of a concept c_j in the feature f. This formula transforms textual domain features into the numeric domain-feature space, informed by the underlying domain ontology. Our design calculates the degree of relevance between an Internet text t (to be classified) and each domain feature f in the domain ontology. These domain relevance indicators provide a high-

level view of the input text, especially regarding domain-specific knowledge.

Ontology Generation Module performs the process of knowledge acquisition and representation, extracting concepts and features from text and representing them on an ontology. In other words, the module is designed to ontologize the identified concepts and relations in each piece of real-time online text and to use the ontology mapping to reduce the dimensionality of feature vectors of natural language texts. For a single concept, we calculate cosine distance to determine its similarity to the existing concept in domain knowledge, $Cosine (a, b) = \sum a_i b_i \Big/ \sqrt{\sum a_i} \sqrt{\sum b_i}$.

Text classification module constructs the feature matrix, incorporating the features in the text ontology and classifies the text concerning its relevance. As our classification task is to determine whether or not the text is relevant (0 for not or 1 for relevant), this study employs a variety of machine learning techniques (e.g., SVM, decision tree, naive Bayes, logistic regression) to regularize and train our model. This study conducts a comparison and evaluation of the results from different machine learning techniques, the best of which is selected into the final instantiation of our proposed design.

6. Design instantiation

This study instantiates our design in the domain of "terrorism." Terrorism, as a difficult research subject, associates a large variety of actors and terrorism activities over time and across the globe [39]. The nature of terrorism has been changing over the past few decades, as more terrorist groups depend less on formal leadership, instead maintaining non-traditional organizational hierarchy [40]. This brings difficulties to researchers and practitioners who are in terrorism observation for further analysis and detection. The majority of existent studies on terrorism are conducted alone, accessing limited data resources from and for non-academicians [39]. Motivated to improve terrorism data discovery and management, the IS community has been making continuous efforts to design IS artifacts in the domain of terrorism [41-46]. While acknowledging the contributions of prior studies, we notice that there is a lack of a dynamic terrorism knowledge discovery system – an SE for terrorism domain, which is intelligent enough to discover and manage terrorism knowledge in a timely method, assisting both practitioners and researchers in discovering and linking the pertinent "dots" to generate useful hints or

warnings [47]. Therefore, we instantiate our design within the domain of "terrorism".

The domain knowledge base in our instantiation is the Global Terrorism Database (GTD) which includes information on global terrorist events from 1970 through 2016 – data on domestic and international terrorist incidents that have occurred, counting up to more than 170,000 cases[1]. Domain-specific features (e.g., weapon, target, attack type, actor, region, country, state, city) that are relevant to "terrorism" are highlighted in the knowledge of GTD, providing the contextual references and serving our calculation of domain features of an input text.

Figure 5. Terrorism domain features in GTD knowledge ontology

Crawling with keyword query – To bypass the anti-crawling mechanism of Google Search, we use Selenium Python. The WebDriver of Selenium enables our crawling module to mimic human behavior of visiting a website, clicking, and collecting the web page elements. The instantiated crawling module accepts keywords as the search query, browsing and identifying the text of interest on the Internet. The example in Figure 6 illustrates the sample data set crawled with the keyword "lone wolf." The search runs continuously until the module receives a termination from the user. As our design focuses on the texts, which are embedded in the web resources of Hypertext Markup Language (HTML) format, we use the text extractor "Goose" (from Gravity natural language processing lab).

_id: ObjectID('59809d56f07170ab341c9929')
title: "Official Merchandise Page of Kid Christian – Pro Wrestling Tees"
link: "http://www.prowrestlingtees.com/wrestler-t-shirts/thatkidkc.html"
id: "2017080111251"
time: "2017-08-01 10:52:10"
description: "T-shirts and more apparel created and sold by Kid Christian"
text: "18 years old and coming up with fire. Kid christian 2 years in the business and becoming the future of pr
ofessional wrestling as we know it. With wins against CWC/205 Live Star Sean Maluta and Big Win against RO
H Top Prospect Lio Rush and becoming a 2x SWF Cruiserweight Champion, KC is proving why he has what it tak
es to make it. Follow him on his road to stardom because he is the Hottest, Most Athletic and Best Kid in
the world Sports Entertainment today...That Kid....KID CHRISTIAN!!!"
title_in_text: "Official Merchandise Page of Kid Christian"
topic: Array

_id: ObjectID('59809d56f07170ab341c992a')
title: "Betting Odds Shift For Tonight's Shinsuke Nakamura vs John Cena ..."
link: "http://www.wrestlezone.com/news/867991-betting-odds-shift-for-tonights-shinsuke-nakamura-vs-john-cena-mat
ch-who-is-favored-to-be-wwe-champion-after-summerslam"
id: "2017080111251"
time: "2017-08-01 10:40:10"
description: "Betting Odds Shift For Tonight's Shinsuke Nakamura vs John Cena Match, Who is Favored to Be WWE Champion
After Summerslam?!"
text: "Here is an interesting prop bet and an update on the odds for tonight's Smackdown match featuring Shinsuk
e Nakamura vs John Cena. There are currently odds on who will be the WWE Champion at the conclusion of thi
s year's Summerslam in Brooklyn. The current champion, Jinder Mahal, has the best odds at +195 and this is
understandable with Mahal being rumored to in for a long championship reign. The other three possibilities
that have odds are Shinsuke Nakamura, John Cena and Baron Corbin. Nakamura at +215 and John Cena at +230 a

Figure 6. Crawling module result sample

Feature extraction – A state-of-the-art neural network-based model is used for coreference resolution [48] in our instantiation. Stanford NER tags and Relation Extractor (OpenIE) are employed for entity extraction and relation extraction. Entities and their relations are extracted from the given text (as is shown in Figure 4). Moreover, terrorism domain-specific features (listed in Figure 5), such as weapon, terrorist organization, attack type, and target, are also identified. The domain relevance scores are calculated by our calculation function implemented in Python.

Figure 7. An example of text ontology

Ontology Generation – To populate the text ontology (Figure 7) with the extracted granular information from the retrieved text, we use JENA framework which provides Java APIs for reading, writing, and processing RDF-based ontologies. In addition, we use Protégé 5.17 to examine OWL ontologies.

Text Classification – Python Scikit-learn and Pandas are used to implement the classification task. A variety of machine learning models are generated

---

[1] https://www.start.umd.edu/gtd/

and compared (e.g., SVM, decision tree, naive Bayes, and logistic regression).

7. Conclusion

In this paper, we put forward a new method for ontology-driven domain-specific SE. In order to handle the limitations with existing methods, our method, leveraging the integrated power of syntax, semantics, and indexical pragmatics, ontologically annotates the text information with respects to the semantic and pragmatic features of the text. The merits of our design include improving the relevance of search results and requiring no large manually annotated training data.

This study has its theoretical contribution. Earlier in this paper, under the section titled "research methodology," we drew upon the comprehensive eight-component framework from Gregor and Jones [30] for articulating an information systems design theory (ISDT). The ISDT components are explicitly explained and our ISDT is expressed in those terms. Another theoretical contribution is our application of the concepts of syntax, semantics, and indexical pragmatics into computational text analysis, illustrating the power of indexical pragmatics in improving SE performance.

This research is highly applicable to practice. Our proposed novel design for domain-specific SE can dynamically retrieve, analyze and determine the relevance of information on the Internet. Additionally, our framework proposes a practical and feasible way through which intelligent systems can leverage the semantic web technique to transform text knowledge into concepts-based ontology concerning linguistic and domain features. Moreover, we appropriately deal with the interface between natural language semantics and pragmatics by leveraging the power of such NLP techniques of coreference resolution and entity and relation extraction. Especially, we illustrate how our proposed design can be used to build up a terrorism domain SE which is of great value to both scholars and practitioners.

While integrating the techniques of NLP and automatic ontology generation to support knowledge discovery and information interoperability in text, our design only allows and targets a specific domain. Another limitation is that the proposed artifact has not been fully evaluated. We plan to assess the utilities of each module in the artifact and to present an evaluation of the operational performance of the artifact. Further, the future research would be expanding the current design with the multi-agent

technique so that it can be deployed for multi-domain use simultaneously.

8. Acknowledgement

9. References

[1] Taylor, A. (2016). 47 percent of the world's population now use the Internet, study says. Retrieved from https://www.washingtonpost.com/news/worldviews/wp/2016/11/22/47-percent-of-the-worlds-population-now-use-the-internet-users-study-says/?utm_term=.c4bbaf79171b

[2] Schmidt, S., Schnitzer, S. and Rensing, C. Text classification based filters for a domain-specific search engine. *Computers in Industry*, 78 (2016), 70-79.

[3] Jain, A. N. Surflex: fully automatic flexible molecular docking using a molecular similarity-based search engine. *Journal of Medicinal Chemistry*, 46, 4 (2003), 499-511.

[4] Widyantoro, D. H. and Yen, J. 2001. A fuzzy ontology-based abstract search engine and its user studies, The 10th IEEE International Conference on Fuzzy Systems: IEEE, pp. 1291-1294.

[5] McCallum, A., Nigam, K., Rennie, J. and Seymore, K. 1999. A machine learning approach to building domain-specific search engines, IJCAI'99 Proceedings of the 16th international joint conference on Artificial intelligence, pp. 662-667.

[6] Croft, W. B., Metzler, D. and Strohman, T. (2010). Search engines: Information retrieval in practice. Boston: Pearson.

[7] Chum, O., Philbin, J., Sivic, J., Isard, M. and Zisserman, A. 2007. Total recall: Automatic query expansion with a generative feature model for object retrieval, The11th IEEE International Conference on Computer Vision: IEEE, pp. 1-8.

[8] Grechanik, M., Fu, C., Xie, Q., McMillan, C., Poshyvanyk, D. and Cumby, C. 2010. A search engine for finding highly relevant applications, The 32nd International Conference on Software Engineering: ACM, pp. 475-484.

[9] Witten, I. H., Nevill-Manning, C., McNab, R. and Cunningham, S. J. A public library based on full-text retrieval. *Communications of the ACM*, 41, 4 (1998), 71-75.

[10] Kamis, A. A. and Stohr, E. A. Parametric search engines: What makes them effective when shopping online for differentiated products? *Information & Management*, 43, 7 (2006), 904-918.

[11] Jones, C. B., Abdelmoty, A. I., Finch, D., Fu, G. and Vaid, S. The SPIRIT spatial search engine: Architecture, ontologies and spatial indexing. *Geographic Information Science* (2004), 125-139.

[12] Tsai, F. S. Web-based geographic search engine for location-aware search in Singapore. *Expert Systems with Applications*, 38, 1 (2011), 1011-1016.

[13] Yang, S.-Y. An ontological website models-supported search agent for web services. *Expert Systems with Applications*, 35, 4 (2008), 2056-2073.

[14] García-Sánchez, F., Martínez-Béjar, R., Contreras, L., Fernández-Breis, J. T. and Castellanos-Nieves, D. An

ontology-based intelligent system for recruitment. *Expert Systems with Applications*, 31, 2 (2006), 248-263.

[15] Bouhana, A., Zidi, A., Fekih, A., Chabchoub, H. and Abed, M. An ontology-based CBR approach for personalized itinerary search systems for sustainable urban freight transport. *Expert Systems with Applications*, 42, 7 (2015), 3724-3741.

[16] Lu, E. J.-L. and Jung, Y.-M. XDSearch: an efficient search engine for XML document schemata. *Expert Systems with Applications*, 24, 2 (2003), 213-224.

[17] Hanbury, A. and Lupu, M. 2013. Toward a model of domain-specific search, The 10th Conference on Open Research Areas in Information Retrieval, pp. 33-36.

[18] Bhogal, J., MacFarlane, A. and Smith, P. A review of ontology based query expansion. *Information Processing & Management*, 43, 4 (2007), 866-886.

[19] Wöber, K. 2006. Journal of Emerging Technologies in Web Intelligence, in Travel Destination Recommendation Systems: Behavioral Foundations and Applications. UK: Biddles Ltd, pp. 205-226.

[20] Finkelstein, L., Gabrilovich, E., Matias, Y., Rivlin, E., Solan, Z., Wolfman, G. and Ruppin, E. 2001. Placing search in context: The concept revisited, The 10th International Conference on World Wide Web: ACM, pp. 406-414.

[21] Gupta, V. and Lehal, G. S. A survey of text mining techniques and applications. *Journal of Emerging Technologies in Web Intelligence*, 1, 1 (2009), 60-76.

[22] d'Aquin, M. and Motta, E. Watson, more than a semantic web search engine. *Semantic Web*, 2, 1 (2011), 55-63.

[23] Xiang, Z., Wöber, K. and Fesenmaier, D. R. Representation of the online tourism domain in search engines. *Journal of Travel Research*, 47, 2 (2008), 137-150.

[24] Geng, B., Yang, L., Xu, C. and Hua, X.-S. Ranking model adaptation for domain-specific search. *IEEE Transactions on Knowledge and Data Engineering*, 24, 4 (2012), 745-758.

[25] Chauhan, R., Goudar, R., Sharma, R. and Chauhan, A. 2013. Domain ontology based semantic search for efficient information retrieval through automatic query expansion, International Conference on Intelligent Systems and Signal Processing (ISSP): IEEE, pp. 397-402.

[26] Hatzi, O., Batistatos, G., Nikolaidou, M. and Anagnostopoulos, D. 2012. A specialized search engine for web service discovery, The 19th IEEE International Conference on Web Services (ICWS): IEEE, pp. 448-455.

[27] Bošnjak, A. and Podgorelec, V. Upgrade of a current research information system with ontologically supported semantic search engine. *Expert Systems with Applications*, 66 (2016), 189-202.

[28] Levinson, S. C. and Levinson, S. C. (2000). Presumptive meanings: The theory of generalized conversational implicature: MIT Press.

[29] Gregor, S. and Hevner, A. R. Positioning and presenting design science research for maximum impact. *MIS Quarterly*, 37, 2 (2013), 337-355.

[30] Gregor, S. and Jones, D. The anatomy of a design theory. *Journal of the Association for Information Systems*, 8, 5 (2007), 312-323,325-335.

[31] Hevner, A., March, S. T., Park, J. and Ram, S. Design science in information systems research. *MIS quarterly*, 28, 1 (2004), 75-105.

[32] Simon, H. A. (1996). The sciences of the artificial. Cambridge, MA: MIT press.

[33] Lee, A. S., Briggs, R. and Dennis, A. R. 2014. Crafting theory to satisfy the requirements of explanation, The 47th Hawaii International Conference on System Sciences (HICSS): IEEE, pp. 4599-4608.

[34] Kuechler, B. and Vaishnavi, V. On theory development in design science research: anatomy of a research project. *European Journal of Information Systems*, 17, 5 (2008), 489-504.

[35] Noy, N. F. Semantic integration: a survey of ontology-based approaches. *ACM Sigmod Record*, 33, 4 (2004), 65-70.

[36] Gruber, T. R. A translation approach to portable ontology specifications. *Knowledge Acquisition*, 5, 2 (1993), 199-220.

[37] Etudo, U., Yoon, V. and Liu, D. Financial Concept Element Mapper (FinCEM) for XBRL interoperability: Utilizing the M3 Plus method. *Decision Support Systems*, 98 (2017), 36-48.

[38] Clark, K. and Manning, C. D. 2015. Entity-centric coreference resolution with model stacking, The 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing, pp. 1405-1415.

[39] Silke, A. (2004). Research on terrorism: Trends, achievements and failures: Routledge.

[40] Ellis, J. O. 2008. Countering terrorism with knowledge, in Terrorism Informatics: Integrated Series in Information Systems. pp. 141-155.

[41] Reid, E. and Chen, H. 2008. Domain mapping of contemporary terrorism research, in Terrorism Informatics: Integrated Series in Information Systems. Springer, pp. 3-26.

[42] Sun, Z., Lim, E.-P., Chang, K., Suryanto, M. A. and Gunaratna, R. K. 2008. Document Selection for Extracting Entity and Relationship Instances of Terrorist Events, in Terrorism Informatics. Springer, pp. 309-346.

[43] Oh, O., Agrawal, M. and Rao, H. R. Information control and terrorism: Tracking the Mumbai terrorist attack through twitter. *Information Systems Frontiers*, 13, 1 (2011), 33-43.

[44] Brynielsson, J., Horndahl, A., Johansson, F., Kaati, L., Mårtenson, C. and Svenson, P. Harvesting and analysis of weak signals for detecting lone wolf terrorists. *Security Informatics*, 2, 1 (2013), 2-11.

[45] Choi, D., Ko, B., Kim, H. and Kim, P. Text analysis for detecting terrorism-related articles on the web. *Journal of Network and Computer Applications*, 38 (2014), 16-21.

[46] Conlon, S. J., Abrahams, A. S. and Simmons, L. L. Terrorism information extraction from online reports. *Journal of Computer Information Systems*, 55, 3 (2015), 20-28.

[47] Popp, R., Armour, T. and Numrych, K. Countering terrorism through information technology. *Communications of the ACM*, 47, 3 (2004), 36-43.

[48] Clark, K. and Manning, C. D. 2016. Deep reinforcement learning for mention-ranking coreference

models, The 2016 Conference on Empirical Methods in Natural Language Processing (EMNLP 2016).