Enabling social web for IoT inducing ontologies from social tagging

Mohammed Alruqimi, Noura Aknin
Information Technology and Modeling Systems Research Unit, Abdelmalek Essaadi University, Morocco

Article Info

Article history:

Received Oct 12, 2018
Revised Dec 20, 2018
Accepted Jan 5, 2019

Keyword:

Internet of things
Social tagging
Social web

ABSTRACT

Semantic domain ontologies are increasingly seen as the key for enabling interoperability across heterogeneous systems and sensor-based applications. The ontologies deployed in these systems and applications are developed by restricted groups of domain experts and not by semantic web experts. Lately, folksonomies are increasingly exploited in developing ontologies. The "collective intelligence", which emerge from collaborative tagging can be seen as an alternative for the current effort at semantic web ontologies. However, the uncontrolled nature of social tagging systems leads to many kinds of noisy annotations, such as misspellings, imprecision and ambiguity. Thus, the construction of formal ontologies from social tagging data remains a real challenge. Most of researches have focused on how to discover relatedness between tags rather than producing ontologies, much less domain ontologies. This paper proposed an algorithm that utilises tags in social tagging systems to automatically generate up-to-date specific-domain ontologies. The evaluation of the algorithm, using a dataset extracted from BibSonomy, demonstrated that the algorithm could effectively learn a domain terminology, and identify more meaningful semantic information for the domain terminology. Furthermore, the proposed algorithm introduced a simple and effective method for disambiguating tags.

Corresponding Author:

Mohammed Alruqimi,
Information Technology and Modeling Systems Research Unit,
Abdelmalek Essaadi University,
Bloc2 App58, Mixta, Martil, Morocco.
Email: m.alruqimi@uae.ma

1. INTRODUCTION

Semantic domain ontologies are increasingly seen as a key factor in automation of information processing. Recently, semantic web technologies are integrating to Internet of Things. These ontologies play an essential role for integrating IoT data and web information systems. Applying such ontologies to IoT would better enable "things" to work in co-operation and also would enable autonomous interaction between "things" [1-6]. However, ontologies development by domain experts is a time-consuming and expensive process. Moreover, the ontologies deployed in the current sensor-based applications are developed by restricted groups of domain experts and not by semantic web experts. In this context, social tagging data contributed by millions of online users represent an essential and continuous source for the "collective intelligence", which are increasingly seen as an alternative to the current effort at semantic web ontologies [7–9]. The ontologies derived from folksonomies can give a machine-processable form of the Social tagging data representing online communities' collective intelligence rather than the perception of a limited group of experts. As such, they would be able to capture changes derived from a more diverse user population. Therefore, they would become semantically richer and thus handier for logical reasoning tasks [10]. Unfortunately, social tagging systems share the problems inherent to all uncontrolled vocabularies, such as ambiguity, synonymy, and the lack of hierarchy. Thus, knowledge extraction from the social tagging data remains a challenge not solved yet. In this paper, we introduce an algorithm for inducing domain ontology from social tagging data. Experimental results, on a snapshot of dataset from BibSonomy, showed that the introduced algorithm could effectively capture domain-specific concepts, and enrich these concepts with semantic information extracted from Wikipedia.

2. SOCIAL TAGGING SYSTEMS

Social tagging websites enable users to assign free-chosen tags to categorize their digital content (such as websites, pictures, videos etc.) over the Web, forming the so-called folksonomies [11]. Currently, many web-based services foster the concept of tagging. These systems can be differentiated according to the kind of resources supported in. For instance, Delicious for sharing bookmarks, Flicker for photos, BibSonomy for publications and bookmarks and YouTube for sharing videos. The basic principle of these services is simply to allow registered users generating the content and classify it in their own unique way by assigning arbitrary tags to this content. Researchers attributed the success of tagging to the fact that no specific prior knowledge is required to tag, and the immediate benefit of tagging [12], [13]. From a knowledge organization point of view, folksonomies have two main advantages: Social tagging systems provide a vast amount number of user-generated annotations and directly reflect users' vocabularies and interests; they are relatively cheap to develop and harvest as they emerge from end users' tagging [12–14]. These advantages have turned Social tagging systems into an interesting data sources for Semantic Web applications [7], [8], [14], [15].

3. RELATED WORKS

Much work has been done to introduce semantics in folksonomy [16-19], and to investigate methods of deploying this semantics for tasks such as information retrieval [20-22], recommender systems [23-26], and ontologies development [27-29]. As well, quite a number of works has been done to extract structured knowledge and develop ontologies from social tagging systems. The early studies explored means of leveraging the co-occurrence statistics of tags and the tripartite structure of folksonomies to measure tag relatedness (e.g., [30-33]). More recently researchers (e.g., [28], [34], [35]) proposed to make tags semantics explicit by grounding them to corresponding entries in online knowledge bases, such as WordNet and DBpedia. Although these approaches are more precision [36], but approaches heavily dependent on WordNet get poor recall due to the fact that many of the tags from folksonomies do not exist in WordNet. In general, there is a lack of methods that extract domain-specific ontologies from folksonomies. Our algorithm produces baseline domain ontologies from tags in folksonomies. The proposed algorithm collects domain-relevant terms from tags relying on a set of domain keywords extracted automatically from Wikipedia pages titles. Then, it identifies the exact meaning of the terms and retrieve semantic information about each term.

4. INDUCING DOMAIN ONTOLOGY

Our algorithm takes the name of a specific domain and a prepared folksonomy dataset as inputs and produces a corresponding domain terminology as output. This algorithm first represents folksonomy resources as an undirected weighted graph. Next, it collects a domain terminology through traversing the resources graph relying on a set of domain keywords extracted automatically from titles of Wikipedia entries. Finally, we extract semantics information about the collected domain terminologies by linking them to their appropriate Wikipedia entries. This includes identifying the intended meaning, attributes and synonyms of the domain terminology. The general method is shown in Figure 1.

4.1. Pre-processing

The pre-processing activity is an important task as it guarantees the quality of the data over which the process is going to be carried out. This activity includes deleting special characters, duplicated tags and prepositions. Furthermore, we used a lexical vector to exclude non-objective tags that caused noisy connections between the resources on the resources graph [17], [23].

4.2. Resources graph generation

A folksonomy can be seen as a tuple A: = (U, T, R), where U, T, and R, are finite sets, whose elements are called users, tags and resources, respectively. Folksonomy can be represented as an undirected tri-partite hyper-graph $G = (V, E)$ where $V = U \cup T \cup R$, is the set of vertices and $E = \{(u, t, r) \mid (u, t, r) \in A\}$ is the set of edges; the tri-partite graph can be folded into two and one-mode graphs [7], [37]. In this work, we adopt this definition using the on-model graph G'= (V', E') in which V' represents the set of resources, and E’ represents the set of weighted edges. Two resources (ri , rj) will be connected if they share at least one tag. In the following section, we describe how to traverse this graph in order to collect the relevant domain terms. To implement this phase, we use JGraphT library, which is a free Java class library that provides mathematical graph-theory objects and algorithms (http://jgrapht.org/).

4.3. Domain Terminology Collection

In this activity, our algorithm starts by extracting a list of Domain Keywords from the titles of Wikipedia articles and redirection pages contained in the main Wikipedia category corresponded to the domain at hand. Based on this Domain Keywords list, we select a set of resources as starting points (SP) for traversing the graph. For a resource, to be marked as starting point, at least two-thirds of the tags assigned to this resource should be found in the Domain Keywords. Next, our algorithm traverses the graph G’ many times (starting from SPs) looking for resources that are relevant to the domain at hand. The tags that are associated to the returned resources will be collected as domain terminologies. In more details, throughout the traversing process, we applied a ranking function over each visited vertex. The ranking function rates the relevance of a vertex to the given domain based on the number and weight of the paths coming from the different seeds to it (See Figure (1), adapted from [28]). Resources that have a ranking value greater than a defined h threshold have been marked as domain-relevant resources, and hence all their associated tags have been gathered as domain-relevant terms. To traverse the graph, we use the breadth first search (BFS) method; once the graph being traversed starting from a particular seed, the traversing process stops whether reaching another seed or reaching a terminal vertex.

$$a(rj)=a'(rj) + \frac{|\{\, t \in T|(u,rj,t)\in A\}\cap\{\, t \in T|(u,ri,t)\in A\}|}{|\{\, t \in T|(u,rj,t)\in A\}|} * \frac{a(ri)}{d} \tag{1}$$

Let us consider ri is the previously visited vertex from which we reachedrj, d is the distance between the current vertex and seed.

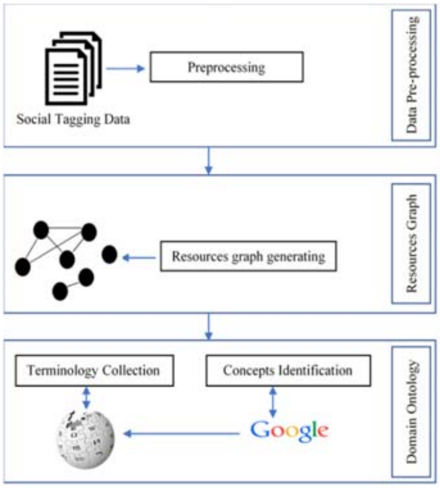

Figure 1. Architecture of the proposed algorithm

4.4. Concepts Identification

By concepts identification, we mean to identify for each term the appropriate Wikipedia article that represents its intended meaning so that we can standardize names of the terms and enrich them by adding their categories and their possible synonyms as well. See the example depicted in Figure 2. This activity also includes disambiguating terms and extracting semantic information about them as well. The advantage of using Wikipedia as a reference to map terms is that Wikipedia is a community-driven knowledge base, much like folksonomies are, so that it rapidly adapts to accommodate new terminology. Many of the popular tags occurring in folksonomies do not appear in grammar dictionaries, such as WordNet, because they correspond to proper nouns, modern technical words, or are widely used acronyms. In addition, the redirect pages in Wikipedia provide synonyms and morphological variations for a concept. For example, when searching the tag ‘nyc’ in Wikipedia, the entry for New York City is returned.

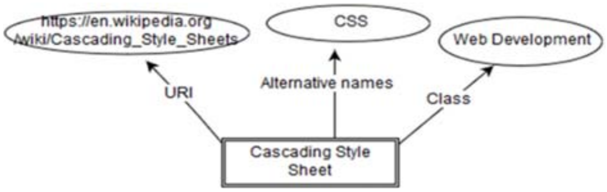

Figure 2. An example of applying Concepts Identification process for the term “CSS”

To perform this task, we used Google as an intermediary to retrieve the appropriate corresponding Wikipedia article for each term. Firstly, we passed to Google a term enclosing between the domain name (in this example: “Web Development”) as a context and the word (“Wikipedia”) to bring Wikipedia pages to the top. Then, we look for a morphological matching between the term and the titles of the top four retrieved Wikipedia pages. The simplest case occurs when a term can be matched directly to the first Google result. In other cases, a term could be matched directly to a page title, to a part of the title, or to one of the redirected pages. As well, terms could be matched to abbreviations that come with the Wikipedia entries’ titles enclosed between parentheses. In some cases, matching to Wikipedia entries fails.

In fact, querying Wikipedia through Google allows taking advantage of techniques embedded in Google, such as stemming and lemmatization, so that we have a high chance of finding the correct corresponding Wikipedia articles. As it shown in Figure 2, passing the term ‘CSS’ to Google resulted in retrieving the Wikipedia article entitled ‘Cascading Style Sheets’ since CSS represents a redirect page to this article in Wikipedia. In the case of disambiguated terms, (for instance the term “Ajax” that could refer to a programming language or a mythological Greek hero), the Wikipedia article that represents its intended meaning comes first in the Google results due to using the domain name as context in Google search box. However, we use information available on the returned Wikipedia articles to enrich the terms. These includes redirect pages as alternative names, and Wikipedia categories containing that page that are listed on the bottom of each article.

5. DISCUSSION

The lack of evaluation frameworks and the lack/incomplete of electronic resources that can be used as a gold standard makes the process of evolution a terminology difficult [17], [38]. Besides, folksonomy tags are uncontrolled vocabularies that contain many slang words and abbreviations, while the electronic resources often use formal and compound terms. However, the experiments were performed on dataset, captured from BibSonomy[39], composed of 20,000 resources annotated by 85,006 tags (11,865 unique tags). Three domains of computer science have been selected randomly for the experiments: Semantic Web, Computer Networks, and Web Development. To evaluate the obtained terminology, we used majority voting of five researchers who were asked to make judgments of domain relevancy (how strongly a term is relevant to the given domain) for all the obtained terms by associating a label “relevant”, “irrelevant”, or “uncertain” with each term. Table 1 shows results we obtained; where the “Distinct Terms” column shows all obtained terms after removing duplicated items, and the “Relevant Terms” column shows the terms marked as domain-relevant terms. We calculated the precision of the obtained results as follows: Precision=(|relevant|) *100 / (|distinct terms|), where distinct terms refer to the all unique terms we obtained. Formally, distinct terms = relevant ∪ irrelevant ∪ uncertain where “relevant” refers to the terms that were marked as domain relevant terms; “irrelevant” refers to the terms that were marked as not domain relevant terms and “uncertain” for unobvious terms.

Table 1. Statistics of the results

Domain	Distinct Terms	Relevant Terms	Precision
Networks	149	62	41.61%
Web Development	165	93	56.36%
Semantic Web	116	77	66.38%

Table 2 shows number of terms that have been correlated successfully to Wikipedia articles. However, we noticed empirically that mapping tags to Wikipedia entries can be used as a good way for excluding non-objective tags since, by nature, the non-objective tags do not have corresponding articles in Wikipedia. Using Google to querying Wikipedia increases the probability of positive matching terms to

Wikipedia entities, as Google can recognize words morphology. Nevertheless, some terms could not be correlated to a Wikipedia article due to missing completed context (e.g. "usability" term cannot be linked, but "web usability" can be). In other cases, terms cannot be matched due to the variant structures of compound terms. For instance, matching the term "DHML" to the article labelled "Dynamic Html" fails although they refer to the same concept. Some terms, such as W3C, XML, could be considered relevant to several domains. As we addressed in early, generating the domain keywords plays an important role for obtaining a good recall. In this context, redirect pages in Wikipedia serve as good domain keywords, as folksonomies contain much neologisms and acronyms. However, for generating another domain concepts by our algorithm, domain experts may be involved in selecting the suitable dataset for a given domain. Users utilize folksonomies with various intentions. For instance, Delicious is used for general purpose whereas BibSonomy primarily serves academic and scientific interests. Compared to general folksonomy, academic folksonomy has a more complex nature in terms of semantics and sparsity of the data [40-42]. Therefore, academic folksonomies would be more useful for building ontologies (particularly, ontologies for scientific domains). Nevertheless, general folksonomies would be more suitable for extracting concepts of general domains such as Movies, Transport. This issue may be considered in our future work. Besides developing a method that looks for corresponding entries on the different online knowledge sources for terms that cannot be mapped to Wikipedia.

Table 2. Number of terms correlated successfully to Wikipedia entries

Domain	Computer Networks	Web Development	Semantic Web
Domain Concepts	43	55	52

6. CONCLUSION

Tag-based systems have become widely available thanks to their advantages, which include self-organization, currency, and ease of use. The bottom-up nature of these systems has proved to be an interesting knowledge source, since they provide a rich terminology generated by potentially large user communities. This paper addressed the problem of how to harvest and exploit embedded semantics in social tagging systems for developing semantic ontologies. The evaluation of the algorithm, using a dataset extracted from BibSonomy, demonstrated that the algorithm could effectively learn domain ontology concepts, and identify meaningful semantic relations for the extracted concepts. Furthermore, the proposed algorithm could help in reducing common problems related to tag ambiguity and synonymous tags.

REFERENCES

[1] L. Atzori, A. Iera, G. Morabito, and M. Nitti, "The Social Internet of Things (SIoT) – When social networks meet the Internet of Things: Concept, architecture and network characterization," *Comput. Networks*, vol. 56, no. 16, pp. 3594–3608, Nov. 2012.

[2] P. Barnaghi, W. Wang, C. Henson, and K. Taylor, "Semantics for the Internet of Things," *Int. J. Semant. Web Inf. Syst.*, vol. 8, no. 1, pp. 1–21, 2012.

[3] A. Gyrard, M. Serrano, and G. A. Atemezing, "Semantic web methodologies, best practices and ontology engineering applied to Internet of Things," in *2015 IEEE 2nd World Forum on Internet of Things (WF-IoT)*, 2015, pp. 412–417.

[4] E. Psomakelis, F. Aisopos, A. Litke, K. Tserpes, M. Kardara, and P. M. Campo, "Big IoT and Social Networking Data for Smart Cities - Algorithmic Improvements on Big Data Analysis in the Context of RADICAL City Applications," in *Proceedings of the 6th International Conference on Cloud Computing and Services Science*, 2016, pp. 396–405.

[5] I. Szilagyi and P. Wira, "Ontologies and Semantic Web for the Internet of Things - a survey," in *IECON 2016 - 42nd Annual Conference of the IEEE Industrial Electronics Society*, 2016, pp. 6949–6954.

[6] N. Lin, F. Tian, E. Sun, and C. Wang, "Swarm Intelligence for Perception Layer Design of Internet of Things," *Inst. Adv. Eng. Sci.*, vol. Vol 12, No, 2014.

[7] T. Gruber, "Ontology of Folksonomy: A Mash-Up of Apples and Oranges," *Int. J. Semant. Web Inf. Syst.*, vol. 3, no. 1, pp. 1–11, 2005.

[8] C. Shirky, "Ontology is Overrated -- Categories, Links, and Tags," 2005. [Online]. Available: http://www.shirky.com/writings/ontology_overrated.html?goback=.gde_1838701_member_179729766. [Accessed: 29-Dec-2016].

[9] I. Peters and W. G. Stock, "Folksonomy and information retrieval," *Proc. Am. Soc. Inf. Sci. Technol.*, vol. 44, no. 1, pp. 1–28, Oct. 2007.

[10] A. Mikroyannidis, "Toward a Social Semantic Web," *Computer (Long. Beach. Calif).*, vol. 40, no. 11, pp. 113–115, Nov. 2007.

[11] T. Vander Wal, "Folksonomy Coinage and Definition." Jun-2007.

[12] A. Mathes, "Folksonomies - Cooperative Classification and Communication Through Shared Metadata," 2004.

[Online]. Available: http://www.adammathes.com/academic/computer-mediated-communication/folksonomies.html. [Accessed: 02-Jul-2016].

[13] A. Hotho, R. Jäschke, C. Schmitz, and G. Stumme, "Information Retrieval in Folksonomies: Search and Ranking," in *Proceedings of the 3rd European conference on The Semantic Web: research and applications*, Springer-Verlag, 2006, pp. 411–426.

[14] M. Szomszor *et al.*, "Folksonomies, the Semantic Web, and Movie Recommendation," 2007.

[15] H. S. Al-Khalifa and H. C. Davis, "Towards better understanding of folksonomic patterns," in *Proceedings of the 18th conference on Hypertext and hypermedia - HT '07*, 2007, p. 163.

[16] A. García-Silva, O. Corcho, H. Alani, and A. Gómez-Pérez, "Review of the state of the art: discovering and associating semantics to tags in folksonomies," *Knowl. Eng. Rev.*, vol. 27, no. 1, pp. 57–85, Mar. 2012.

[17] M. Alruqimi and N. Aknin, "Semantic Emergence From Social Tagging Systems," *Int. J. Organ. Collect. Intell.*, vol. 5, no. 1, pp. 16–31, 2015.

[18] F. Jabeen, S. Khusro, A. Majid, and A. Rauf, "Semantics discovery in social tagging systems: A review," *Multimed. Tools Appl.*, vol. 75, no. 1, pp. 573–605, Jan. 2016.

[19] H. Liu, H. Chen, M. Lin, and Y. Wu, "Community Detection Based on Topic Distance in Social Tagging Networks," *TELKOMNIKA Indones. J. Electr. Eng.*, vol. 12, no. 5, May 2014.

[20] M. N. Uddin, T. H. Duong, N. T. Nguyen, X. M. Qi, and G. S. Jo, "Semantic similarity measures for enhancing information retrieval in folksonomies," *Expert Syst. Appl.*, 2013.

[21] A. Zubiaga, V. Fresno, R. Martinez, and A. P. Garcia-Plaza, "Harnessing Folksonomies to Produce a Social Classification of Resources," *IEEE Trans. Knowl. Data Eng.*, vol. 25, no. 8, pp. 1801–1813, Aug. 2013.

[22] A. Tommasel and D. Godoy, "Semantic grounding of social annotations for enhancing resource classification in folksonomies," *J. Intell. Inf. Syst.*, vol. 44, no. 3, pp. 415–446, Jun. 2015.

[23] I. Cantador, I. Konstas, and J. M. Jose, "Categorising social tags to improve folksonomy-based recommendations," *Web Semant. Sci. Serv. Agents World Wide Web*, vol. 9, no. 1, pp. 1–15, Mar. 2011.

[24] I. Ching Hsu, "Integrating ontology technology with folksonomies for personalized social tag recommendation," *Appl. Soft Comput.*, vol. 13, no. 8, pp. 3745–3750, Aug. 2013.

[25] F. Font, J. Serrà, and X. Serra, "Analysis of the Impact of a Tag Recommendation System in a Real-World Folksonomy," *ACM Trans. Intell. Syst. Technol.*, 2015.

[26] A. M. El-korany and S. M. Khatab, "Ontology-based Social Recommender System," *Inst. Adv. Eng. Sci.*, vol. Vol 1, no. 3, 2012.

[27] S. Hamdi, A. Lopes Gancarski, A. Bouzeghoub, and S. Ben Yahia, "Enriching Ontologies from Folksonomies for eLearning: DBpedia Case," in *2012 IEEE 12th International Conference on Advanced Learning Technologies*, 2012, pp. 293–297.

[28] A. García-Silva, L. J. García-Castro, A. García, and O. Corcho, "Social Tags and Linked Data for Ontology Development: A Case Study in the Financial Domain," in *Proceedings of the 4th International} Conference on Web Intelligence, Mining and Semantics (WIMS14)*, 2014, p. 32:1--32:10.

[29] S. Wang, W. Wang, Y. Zhuang, and X. Fei, "An ontology evolution method based on folksonomy," *J. Appl. Res. Technol.*, vol. 13, no. 2, pp. 177–187, Apr. 2015.

[30] G. Begelman, P. Keller, and F. Smadja, "Automated Tag Clustering: Improving searching and exploration in the tag space," in *WWW2006*, 2006.

[31] P. Schmitz, "Inducing Ontology from Flickr Tags," in *Proceedings of the Workshop on Collaborative Tagging at WWW2006*, 2006.

[32] P. Heymann and H. Garcia-Molina, "Collaborative Creation of Communal Hierarchical Taxonomies in Social Tagging Systems." 2006.

[33] P. Mika, "Ontologies Are Us: A Unified Model of Social Networks and Semantics," Springer Berlin Heidelberg, 2005, pp. 522–536.

[34] S. Angeletou, "Semantic Enrichment of Folksonomy Tagspaces," in *The Semantic Web - ISWC 2008*, Berlin, Heidelberg: Springer Berlin Heidelberg, 2008, pp. 889–894.

[35] I. Cantador, M. Szomszor, H. Alani, M. Fernandez, and P. Castells, "Enriching ontological user profiles with tagging history for multi-domain recommendations," in *1st International Workshop on Collective Semantics: Collective Intelligence &amp; the Semantic Web (CISWeb 2008)*, 2008.

[36] J. Wei and F. Meng, "Is collective intelligence helps more in polysemy tag optimazed algorithm than commonsense tool," *Cluster Comput.*, Jul. 2017.

[37] P. Mika, "Ontologies Are Us: A Unified Model of Social Networks and Semantics," Springer Berlin Heidelberg, 2005, pp. 522–536.

[38] K. Dellschaft and S. Staab, "On How to Perform a Gold Standard Based Evaluation of Ontology Learning," in *Proceedings of the 5th international conference on The Semantic Web*, Springer-Verlag, 2006, pp. 228–241.

[39] "Knowledge & Data Engineering Group, University of Kassel: Benchmark Folksonomy Data from BibSonomy, version of September 30st, 2008." [Online]. Available: https://www.kde.cs.uni-kassel.de/bibsonomy/dumps/.

[40] H. Du, S. K. W. Chu, and F. T. Y. Lam, "Social bookmarking and tagging behavior: an empirical analysis on delicious and connotea," in *Proceedings of the 2009 International Conference on Knowledge Management*, 2009.

[41] D. H. Lee, "Comparative Analysis of Index Terms and Social Tags," *J. KOREAN Soc. Libr. Inf. Sci.*, vol. 49, pp. 291–311, 2015.

[42] H. Dong, W. Wang, and F. Coenen, "Deriving Dynamic Knowledge from Academic Social Tagging Data: A Novel Research Direction," in *iConference 2017 Proceedings*, 2017, pp. 661–666.