Generating Computational Taxonomy for Business Models of the Digital Economy

Chao Wu[1(✉)], Yi Cai[2], Mei Zhao[2], Songping Huang[2], and Yike Guo[1]

[1] Data Science Institute, Imperial College London, London, UK
{chao.wu,y.guo}@imperial.ac.uk
[2] School of Software Engineering,
South China University of Technology, Guangzhou, China
ycai@scult.edu.cn, {1527757976,798623154}@qq.com

Abstract. We propose to design a semi-automatic ontology building approach to create a new taxonomy of the digital economy based on a big data approach – harvesting data by scraping publicly available Web pages of digitally-focused business. The method is based on a small core ontology which provides the basic level concepts in business model. We try to use computational approaches to extracting Web data towards generating concepts and taxonomy of business models in the digital economy, which can help consequently address the important question while exploring new business models in big data era.

Keywords: Business model taxonomy · Ontology generation · Computational taxonomy

1 Introduction

Big Data and data science promises to change how business is conducted and how innovations emerge. Publicly accessible data, such as Web pages, provide a rich source of structured and unstructured data that we can begin to study and extract knowledge about emerging trends in culture, society and business – and about the digital economy itself.

In the digital economy many new companies emulate their business models, for example Google's exploitation of Web user traffic as a means to generating advertising revenue is commonly mimicked. Many companies also innovate new value streams. The ways in which value is generated in the digital economy are not just unclear, but are also still emerging. Value configurations are very different in digital business, where companies frequently partner with other digital revenue streams. Current business model taxonomies fail to reflect this dynamism.

Taxonomy are data schemas, providing a controlled vocabulary of concepts, each with an explicitly defined and machine processable semantics. By defining shared and common domain theories, taxonomy help both people and machines to communicate concisely, supporting the exchange of semantics and not only syntax. The main problems are how to construct domain-specific taxonomy cheaply and quickly. Still now, the generations of most of the taxonomy for business models depend on human.

© Springer International Publishing Switzerland 2016
H. Gao et al. (Eds.): DASFAA 2016 Workshops, LNCS 9645, pp. 126–133, 2016.
DOI: 10.1007/978-3-319-32055-7_11

The semi-automatic and automatic generation method is far from sophisticated and practical. The manual acquisition of ontologies still remains a tedious, cumbersome task.

What we propose is to address this challenge to designing a semi-automatic ontology building approach which creates a new taxonomy of the digital economy by taking a big data approach – harvesting data by scraping publicly available Web pages of digitally-focused business, and processing this data using text analytics techniques including text feature extraction, natural language processing, and supervised learning. This approach is based on a small core ontology which provides the basic level concepts in business model. Cognitive psychologists find that most human knowledge is represented by basic level concepts which is a family of concepts frequently used by people in daily life. In this work, based on a small core ontology constructed by domain experts, we try to use computational approaches to extracting Web data towards generating concepts and taxonomy of business models in the digital economy, which can help consequently address the research question, "What are the new business models of the digital economy?".

The remainder of this paper is organized as follows. Section 2 overviews of related works. Section 3 presents our approach for taxonomy learning method. Section 4 then presents the experiments and evaluations. Finally, Sect. 5 concludes the paper.

2 Related Work

Taxonomic classification of Web pages by computational means is not novel [2], and text mining has been used extensively in applied sciences, for example in biology to extract linked medical concepts, map biological processes, and even to aid the interpretation of genes for drug development.

"Taxonomy" or broadly "Ontology" in its original sense is a philosophical discipline dealing with the potentialities and conditions of being. Within Computer Science, 'ontologies' have been introduced about a decade ago as a means for formally representing knowledge. Gruber gave out the most popular definition of ontology an "explicit, specification of a conceptualization" [6]. This means that ontologies serve as representation in some pre-defined formalism of those concepts and their relations that are needed to model a certain application domain.

Ontology learning can be defined as the set of methods and techniques used for building ontology from scratch, enriching, or adapting an existing ontology in a semi-automatic fashion using several sources. Several approaches exist for the partial automatization of the knowledge acquisition process. To carry out this automatisation, natural language analysis and machine learning techniques can be used.

Maedche and Staab [7] distinguished different ontology learning approaches focus on the type of input used for learning, such as semi-structured text, structured text, unstructured text. In this sense, they proposed the following classification: ontology learning from text, from dictionary, from knowledge base, from semi-structured schema and from relational schema. Now, most of the domains haven't so much existed semi-structured text, structured text, but there are many unstructured text, such as domain literature, web page. So most of the method is to learn ontology from texts consist of

extracting ontologies by applying natural language analysis techniques to texts. The most well-known approaches from this groups are:

1. Ontology pruning is to build a domain ontology based on different heterogeneous sources. It has the following steps. First, a generic core ontology is used as a top level structure for the domain-specific ontology. Second, a dictionary is used to acquire domain concepts. Third, domain-specific and general corpora of texts are used to remove concepts that were not domain specific. This method can quickly construct aim ontology for a specific domain, but for the lack of domain generic core ontology and the efficient method of pruning still now, the effect of exist application is not so good [8].
2. Conceptual clustering, concepts are grouped according to the semantic distance between each other to make up hierarchies. But because of lack the domain context to instruct in the process of distance computation, the conceptual clustering process can't be efficiently controlled. Furthermore, by this method, only taxonomic relations of the concepts in the ontology can be generated [9].
3. Formal concept analysis (FCA), by some technique of NLP, the domain concepts and their attributes can be obtained to form the formal context for the construction of concept lattice. This concept lattice can be viewed as original ontology which just contains classification relations between concepts. After adding non-taxonomic relations, the ontology can be formed. But this difficulty of this method is concept lattice is complicated data structure, when formal context is big, the ontology construction from a set of relevant documents where construction of concept lattice is not easy.
4. Association rules, the association rules have been used to discover non-taxonomic relations between concepts, using a concept hierarchy as background knowledge. Association rules are most used on the data mining process to discover information stored on database. Ontology learning mostly uses unstructured texts but not the structure data in database. So, association rule is just an assistant method to help the ontology generation [7].
5. Pattern-based extraction, a relation is recognized when a sequence of words in the text matches a pattern. But the pattern should be created under the domain expert's instruction. The modification of pattern will bring vibration effect and there is no promise of best pattern [10].
6. Concept learning, a given taxonomy is incrementally updated as new concepts are acquired from real-world texts. Concept learning is a part of the process of ontology learning [11].

The main lacks for all the methods and tools presented in this overview are that there are not integrated methods and tools that combine different learning techniques and heterogeneous knowledge sources with existing ontologies to accelerate the learning process.

Research into taxonomies of e-business has been previously carried out [3, 4], however our approach looks to keep pace with the quickly evolving nature of businesses in today's digital economy by using computational techniques to steer such classification.

Our work is based on a core set of basic concepts. According to the studies of cognitive psychology, there is a family of categories named basic level categories [12, 13]. People most frequently prefer to use basic level concepts constructed from these categories in their daily life, and these concepts are the ones first named and understood by children. For example, when people see a dog, although we also can call it an "animal" or a "terrier", most people would call it a "dog". What is more, most human knowledge is represented by basic level concepts.

3 Method

The approach starts from a core set of concepts built by Human experts, which provides the system with a small number of domain-specific top concepts that represent high-level concepts and are used as seed concepts to discover new concepts and relations [14]. Those concepts and their relations are viewed as the core ontology of the system. Ordinarily, the domain's name can be the top concept of the core ontology. First, the documents in domain corpus need to be preprocessed and converted into plain text format that natural language process tools can conduct. Then, the natural language process tools including stemming, pos tagging and parsing tools used to process the plain texts. Using Gate 3.0 [20] as the tools which is general architecture for text engineering and open source tool, the stemming and pos tagging results of the text will be obtained and stored to the semantic units database.

Then, each sentence of the text will be sent into Stanford Parser [15], which works out the grammatical structure of sentences and the parser tree of each sentence will be analyzed to produce semantic units. The semantic unit is the minimal sentence segment that can be independently viewed as a sentence in a sentence. Then, each of the previous semantic units were POS-tagged and parsed. The process of the corpus processing is shown in Fig. 1.

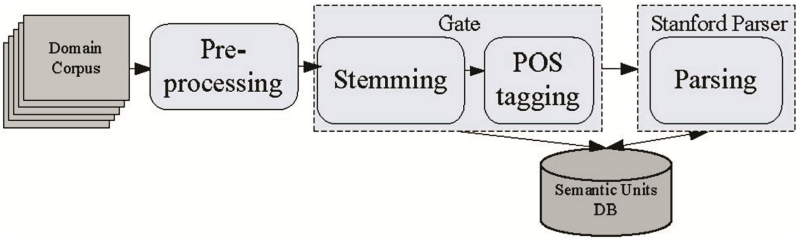

Fig. 1. Domain corpus processing process

As described in [5], the core domain concept or corpus is defined as a tuple $F := (U, T, R, Y)$ where U, T and R are finite sets, whose elements are called users, tags and resources, respectively, and Y is a ternary relation over them, i.e. $Y \subseteq U \times T \times R$.

The target is to learn a hierarchical taxonomy of concepts, which is a tuple $O = (C, P, I, S)$ where C, P and I are finite sets, whose elements are called concepts, properties and instances, respectively, and S is a set of rules, propositions or axioms that

specify the relations among concepts, properties and instances. Every concept consists of a category of instances and is described by its properties.

Accordingly, an instance is represented as a vector of tag-value pairs: An instance, r_i, is represented by a vector of tag:value pairs, $r_i = (t_{i,1} : v_{i,1}, t_{i,2} : v_{i,2},\ldots, t_{i,n} : v_{i,n})$ with $t_{i,k} \in T$, $0 < v_{i,k} \leq 1$, $1 \leq k \leq n$. Where n is the number of the unique tags assigned to resource r_i, $v_{i,k}$ is the weight of tag $t_{i,k}$ in resource r_i. The weight $v_{i,k}$ determines the importance of the tag $t_{i,k}$ to resource r_i. We consider that a tag assigned by more users to a resource is more important because more users think the tag is useful to describe the resource.

A concept is the abstraction of a category of instances and holds the common properties of them: *A concept, c_i, is represented by a vector of tag:value pairs, $c_i = (t_i, 1 : v_{i,1}, t_{i,2} : v_{i,2},\ldots, t_{i,n} : v_{i,n})$ with $t_{i,k} \in T$, $0 < v_{i,k} \leq 1$, $1 \leq k \leq n$.* Where n is the number of unique tags, $t_{i,k}$ is a common tag of a category of resources, $v_{i,k}$ is the weight of the tag $t_{i,k}$.

Accordingly, we construct a concept through extracting common tags of a category of instances. These common tags are considered as the properties of the concept. The weights of these tags are their mean values among all instances in a category.

Given a set C of categories and a set F of features, the category utility is defined as follows:

$$cu(C, F) = \frac{1}{m} \sum_{k=1}^{m} p(c_k) \left[\sum_{i=1}^{n} p(f_i|c_k)^2 - \sum_{i=1}^{n} p(f_i)^2 \right]$$

where $p(f_i|c_k)$ is the probability that a member of category c_k has the feature f_i, $p(c_k)$ is the probability that an instance belongs to category c_k, $p(f_i)$ is the probability that an instance has feature f_i, n is the total number of features, m is the total number of categories. Features of instances are represented by tags in folksonomies. Accordingly, in the definition of category utility, the tag set T is used as the feature set F and a tag t_i is used as a feature f_i. As we model, the importance of tags is different in folksonomies. To take the differences of tag importance into account, we modify the definition and add the weight w_i of tag t_i into the definition:

$$cu(C, T) = \frac{1}{m} \sum_{k=1}^{m} p(c_k) \left[\frac{\sum_{i=1}^{n_k} w_i p(t_i|c_k)^2}{n_k} - \frac{\sum_{i=1}^{n} w_i p(t_i)^2}{n} \right]$$

To reflect the mean weight of a tag, w_i is defined as $w_i = \frac{1}{N_{ti}} \sum_{j=1}^{N_{ti}} v_{j,i}$, where N_{ti} is the number of resources annotated as the weighted category utility.

Because basic level categories (and concepts) have the highest category utility, the problem of finding basic level categories (and concepts) becomes an optimization problem using category utility as the objective function. The value of category utility is influenced by the intra-category similarity which reflects the similarity among members of a category. Categories with higher intra-category similarity have higher value of category utility. Accordingly, we put the most similar instances together in every step of our method until the decrease of category utility. To compute the similarity, we use

the idf-cosine coefficient [16] which is a commonly used method of computing similarity between two vectors in information retrieval. It is defined as follows:

$$sim(a,b) = \frac{\sum_{k=1}^{n} idf(t_k) \cdot v_{a,k} \cdot v_{b,k}}{\sqrt{\sum_{k=1}^{n} v_{a,k}^2} \cdot \sqrt{\sum_{k=1}^{n} v_{b,k}^2}}$$

where a, b are two concepts, n is the total number of unique tags describing them, and $v_{a,k}$ is the value of tag $t_{a,k}$ in concept a, if a does not have the tag, the value is 0. $idf(t_k)$ is the inverse document frequency of the tag t_k, $idf(t_k) = \log_N(N/N_{tk})$, where N is the total number of resources and N_{tk} is the number of resources annotated by tag t_k, $0 \leq idf(t_k) \leq 1$. When $idf(t_k)$ is 0, the tag t_k is assigned to all resources. In this case, all resources have this tag, the tag is not useful for categorization.

In our algorithm, firstly, we consider every single instance itself as a concept. This type of concept which only includes one instance is considered as the bottom level concepts. Secondly, we compute the similarity between each pair of concepts and build the similarity matrix. Thirdly, the most similar pair in the matrix is identified and merged into a new concept. The new concept contains all instances of the two old concepts and holds their common properties. After that we reconsider the similarity matrix of the remaining concepts. We apply this merging process until only one concept is left or the similarity between the most similar concepts is 0. We then determine the step where the categories have the highest category utility which is the local optimum of category utility. These categories are considered as the basic level categories and the concepts are considered as the basic level concepts. The time complexity is $O(N^2 \log N)$ where N is the number of resources.

To build the taxonomy, we first generate a root concept including all instances. After finding the basic level concepts with Algorithm 1, we add the basic level concepts to the taxonomy as sub-concepts of the root. After several iterations, a cognitively basic ontology is built. The psychological character differentiates the ontology built through our method to the ontology built using methods proposed in previous taxonomy learning research.

4 Result

With the method describe before, we construct business model taxonomy as a pilot study. The construction process starts from core business model ontology with top concepts containing 300 articles from 20 business model related website. Each articles have no more than 5000 words. After once extension by our method, there are respectively over 400 of four types of concepts and relations.

Experiments show that the ontologies generated using our method are more consistent with human thinking. Figure 2 gives an example of the ontology explored through our approach. In our approach, concepts are represented by the common tags of a category of resources. The tags of a concept are inherited by its sub-concepts and a concept has all instances of its descendants. For example, tags “crowdfunding” represents a concept within the category of “crowdsourcing”. Such a representation can keep more

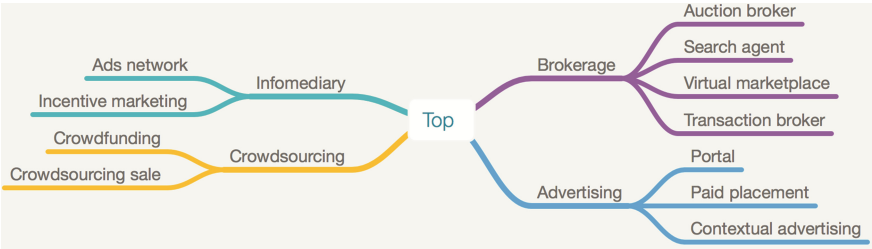

Fig. 2. An ontology generated by our approach.

information and properties of concepts and is consistent with definition of concepts in psychology.

5 Conclusion

In this work, based on a small core ontology constructed by domain experts, we used computational approaches to extracting Web data towards generating concepts and taxonomy of business models in the digital economy. To the best of our knowledge, it is the first work on discovering taxonomy from core concepts for business model. With the result graph-based taxonomy, we can explore the interconnectedness within the digital economy. For example, visual analytics, in the form of interactive graph visualizations, can then be used to explore complex relationships between digital businesses, such as when digital revenue streams cross and are shared within a business model.

Acknowledgment. The work presented in this paper is supported by NEMODE Network + Pilot Study: A Computational Taxonomy of Business Models of the Digital Economy (P55805).

References

1. Mitchell, G.: How many petabytes is the internet? Focus Science and Technology Magazine (2013). http://sciencefocus.com/qa/how-many-terabytes-data-are-internet
2. Xiaoguang, Q., Davison, B.D.: Web page classification: features and algorithms. ACM Comput. Surv. **41**(2) (2009). Article 12
3. Bartelt, A., Lamersdorf, W.: A multi-criteria taxonomy of business models in electronic commerce. In: Fiege, L., Mühl, G., Wilhelm, U.G. (eds.) WELCOM 2001. LNCS, vol. 2232, pp. 193–205. Springer, Heidelberg (2001)
4. Lambert, S.C.: Do we need a 'real' taxonomy of e-business models? School of commerce research paper series 06-06 (2006)
5. Chen, W.H., Cai, Y., Leung, H.F., Li, Q.: Generating ontologies with basic level concepts from folksonomies. Procedia Comput. Sci. **1**(1), 573–581 (2010)
6. Gruber, T.: A translation approach to portable ontology specifications. Knowl. Acquisition **5**(2), 199–220 (1993)
7. Maedche, A., Staab, S.: Ontology learning for the semantic web. IEEE Intell. Syst. **16**(2), 72–79 (2001). Special Issue on the Semantic Web

8. Kietz, J.U., Maedche, A., Volz, R.: A method for semi-automatic ontology acquisition from a corporate intranet. In: Aussenac-Gilles, N., Biebow, B., Szulman, S. (eds.) CEUR Workshop Proceedings, EKAW 2000 Workshop on Ontologies and Texts, Juan-Les-Pins, France, Amsterdam, The Netherlands, vol. 51, pp. 4.1–4.14 (2000). http://CEUR-WS.org/Vol-51
9. Faure, D., Poibeau, T.: First experiments of using semantic knowledge learned by ASIUM for information extraction task using INTEX. In: Staab, S., Maedche, A., Nedellec, C., Wiemer-Hastings, P. (eds.) Proceedings of the Workshop on Ontology Learning, 14th European Conference on Artificial Intelligence ECAI 2000, Berlin, Germany (2000)
10. Morin, E.: Automatic acquisition of semantic relations between terms from technical corpora. In: Proceedings of the Fifth International Congress on Terminology and Knowledge Engineering (TKE 1999). TermNet-Verlag, Vienna (1999)
11. Hahn, U., Schulz, S.: Towards very large terminological knowledge bases: a case study from medicine. In: Hamilton, H.J. (ed.) Canadian AI 2000. LNCS (LNAI), vol. 1822, pp. 176–186. Springer, Heidelberg (2000)
12. Rosch, E., Mervis, C.B., Gray, W.D., Johnson, D.M., Boyes-Braem, P.: Basic objects in natural categories. Cogn. Psychol. **8**(3), 382–439 (1976)
13. Murphy, G.: The Big Book of Concepts. Bradford Book, Cambridge (2004)
14. Zhou, W., Liu, Z., Zhao, Y., Xu, L., Chen, G., Wu, Q., Qiang, Y.: A semi-automatic ontology learning based on wordnet and event-based natural language processing. In: 2006 International Conference on Information and Automation, ICIA 2006, pp. 240–244. IEEE, December 2006
15. The Stanford Parser. http://nlp.stanford.edu/software/lex-parser.shtml
16. Schultz, J., Liberman, M.: Topic detection and tracking using idf-weighted cosine coefficient. In: Broadcast News Workshop 1999 Proceedings, p. 189 (1999)