Incorporating Structured World Knowledge into Unstructured Documents via Heterogeneous Information Networks

Yangqiu Song
Department of Computer Science and Engineering
Hong Kong University of Science and Technology
yqsong@cse.ust.hk

Abstract

Machine learning algorithms have been widely used in document categorization, including both classification and clustering. Two major problems of machine learning in practice are how to generate or extract features from data and how to acquire enough labels for machines to learn. There have been many studies about feature engineering and labeling work reduction in the past decades.

Document feature representation and semantic similarity/relatedness. Document similarity is a fundamental task, and can be used in many applications such as document classification, clustering and ranking. Traditional approaches use bag-of-words (BOW) as document representation and compute the document similarities using different measures such as cosine, Jaccard, and dice. However, the entity phrases rather than just words in documents can be critical for evaluating the relatedness between texts. For example, “New York” and “New York Times” represent different meanings. “George Washington” and “Washington” are similar if they both refer to person, but can be rather different otherwise. Moreover, the links between entities or words are also informative. If we can build a link between “Obama” of type Politician in one document and “Bush” of type Politician in another, then the two documents become similar in the sense that they both talk about politicians and connect to “United States.” Therefore, we can use the structural information in the unstructured documents to further improve document similarity computation.

There have been other approaches trying to incorporate contextual information based on the co-occurrence of words in contexts to solve the above problems. For example, topic model such as Latent Dirichlet Allocation [Blei et al., 2003] uses a generative model to model documents. It assumes that a document is a mixture of topics, where each topic is a mixture of words. Neural network language models such as word2ve [Mikolov et al., 2013] assumes the words can predict its contextual words or can be predicted by its contextual words in a small window. The co-occurrence of words, either in a document or in a window of a central word, can reveal the semantic similarity or relatedness of words and further documents. However, such models yet cannot recover the higher order information. Moreover, the entity types in the higher order relation can be useful to discriminate subtle differences. For example, two documents talking about basketball can be related to different events, as the following two meth-path show:

Doc. → Basketball → NBA ← Basketball ← Doc.
Doc. → Basketball → Olympics ← Basketball ← Doc.

Labeling work reduction. Labeling large amount of data for each domain-specific problems can be very time consuming and costly, especially when there are a lot domains related to different label spaces. Machine learning community has also elaborated to reduce the labeling work done by human for supervised machine learning algorithms or to improve unsupervised learning with only minimum supervision. For example, semi-supervised learning [Chapelle et al., 2006] is proposed to use only partially labeled data and a lot of unlabeled data to perform learning. Transfer learning [Pan and Yang, 2010] uses the labeled data from other relevant domains to help the learning task in the target domain. Both semi-supervised learning and transfer learning need domain knowledge, and there are multiple ways to achieve these learning settings. However, there is no general solution or a principle when applying both learning settings to most tasks. In other words, for each of the target domain, specific domain knowledge is still needed to be engineered into the learning process. Crowdsourcing [Lease, 2011] has also been considered to acquire cheap labels from general-level human intelligence. However, current crowdsourcing mechanisms can still be applied to relatively simple and well-defined tasks, and it is still a challenge for applying machine learning to the labels for more diverse and more specific data [Lease, 2011].

Therefore, we want to seek more general ways to apply the existing knowledge to different domains. Fortunately, with the proliferation of general-purpose knowledge bases (or knowledge graphs), e.g., Cyc project [Lenat and Guha, 1989], Wikipedia, Freebase [Bollacker et al., 2008], Know-ItAll [Etzioni et al., 2004], TextRunner [Banko et al., 2007], WikiTaxonomy [Ponzetto and Strube, 2007], Probbase [Wu et al., 2012], DBpedia [Auer et al., 2007], YAGO [Suchanek et al., 2007], NELL [Mitchell et al., 2015] and Knowledge Vault [Dong et al., 2014], we have an abundance of available world knowledge. We call these knowledge bases world knowledge [Gabrilovich and Markovitch, 2005], because they are universal knowledge that are either collaboratively annotated by human labelers or automatically extracted from big data. When world knowledge is annotated or ex-
tracted, it is not collected for any specific domain. However, because we believe the facts in world knowledge bases are very useful and of high quality, we propose using them as supervision for many machine learning problems. Previously, machine learning algorithms using world knowledge just treat world knowledge as “flat” features in addition to the original text data [Gabrilovich and Markovitch, 2009; Chang et al., 2008; Song and Roth, 2014; Song and Roth, 2015; Song et al., 2016; Song et al., 2011; Song et al., 2015b; Song et al., 2015a]. People have found it useful to use world knowledge as distant supervision for entity and relation extraction [Mintz et al., 2009]. This is a direct use of the facts in world knowledge bases, where the entities in the knowledge bases are matched in the context regardless the ambiguity. A more interesting question is can we use the world knowledge to “supervise” more machine learning algorithms or applications? Particularly, if we can use world knowledge as indirect supervision, then we can extend the knowledge about entities and relations to more generic text analytics problems, e.g., categorization and information retrieval. Song et al. [Song et al., 2013] considered using fully unsupervised method to generate constraints of words using an external general-purpose knowledge base, WordNet. This can be regarded as an attempt to use general knowledge as indirect supervision to help clustering. However, the knowledge from WordNet is mostly linguistically related. It lacks of the information about named entities and their types. Moreover, their approach is still a simple application of constrained co-clustering, where it misses the rich structural information in the knowledge base.

Recently we have developed how to incorporate structured world knowledge into unstructured document categorization via heterogeneous information networks. A heterogeneous information network (HIN) is defined as a graph of multi-typed entities and relations [Han et al., 2010]. Different from traditional graphs, HIN incorporates the type information which can be useful to identify the semantic meaning of the paths in the graph [Sun et al., 2011]. Original HINs are developed for the applications of scientific publication network analysis [Sun et al., 2011; Sun et al., 2012]. Then social network analysis also leverages this representation for user similarity and link prediction [Kong et al., 2013; Zhang et al., 2013; Zhang et al., 2014]. Seamlessly, we can see that the knowledge in world knowledge bases, e.g., Wikipedia, Freebase [Bollacker et al., 2008], YAGO [Suchanek et al., 2007], and NELL [Mitchell et al., 2015], can be represented as an HIN, since the entities and relations in the knowledge base are all typed.

To solve the above challenges of representation and labeling, we investigated in the following studies of incorporating world knowledge into document categorization.

1. We can convert an unstructured textual document into structured heterogeneous information network [Wang et al., 2015a; Wang et al., 2016b]. We proposed the unsupervised semantic parsing based on the previous system [Berant and Liang, 2014] using λ-DCS [Liang, 2013], and proposed three global frequency based inference techniques to rank the logical forms, where the resulting best one is based on the conceptualization framework [Song et al., 2011; Song et al., 2015b].

2. We proposed to extend the meta-path based similarity [Sun et al., 2011] as an ensemble of hundreds of meta-path based similarities [Wang et al., 2015b], where we tried two different unsupervised meta-path ranking algorithm following [Song et al., 2009] to rank different meta-paths, and applied the similarity to both spectral clustering [Wang et al., 2015b] and SVM based classification [Wang et al., 2016a].

3. We proposed to convert a document clustering problem into an HIN partition problem [Wang et al., 2015a; Wang et al., 2016b], and extended information theoretic co-clustering [Dhillon et al., 2003] and constrained information theoretic co-clustering [Song et al., 2013] to propagate the entity type information as constraints to the document cluster assignment decision. The entity type based constraints are considered as indirect supervision for the document clustering problem.

4. We also have some related work on how to cluster documents based on relation expressions as constraints [Wang et al., 2015c] in addition to entity type constraints, relation retrieval from knowledge base [Wang et al., 2016c].

In the talk, I will introduce how we convert the unstructured textual documents into heterogeneous information network representation. Then I will introduce how to compute all the meta-paths efficiently, and compute the weights of different meta-paths for the ensemble of meta-path similarity. Moreover, I will show the results of spectral clustering and SVM classification based on the developed similarity on two benchmark datasets, i.e., 20-newsgroups [Lang, 1995] and RCV1 [Lewis et al., 2004]. I will also introduce how to perform indirect supervision using the extended constrained information theoretical co-clustering algorithm. Finally I will conclude my talk and look forward to having more discussion about current state and future work with the related challenges and problems.

Acknowledgment

The author would thank the collaborators for the support and suggestions about this talk. The author would also thank the workshop organizers for the kind invitation and support.

References


[Han et al., 2010] Jiawei Han, Yizhou Sun, Xifeng Yan, and Philip S. Yu. Mining knowledge from databases: An information network analysis approach. In SIGMOD, pages 1251–1252, 2010.

[Han et al., 2016] Tao Han, Hailong Sun, Yangqiu Song, Yili Fang, and Xu dong Liu. Incorporating external knowledge into crowd intelligence for more specific knowledge acquisition. In IJCAI, 2016.


[Sun et al., 2011] Yizhou Sun, Jiawei Han, Xifeng Yan, Philip S. Yu, and Tianyi Wu. Pathsim: Meta path-based top-k similarity search in heterogeneous information networks. PVLDB, pages 992–1003, 2011.

[Sun et al., 2012] Yizhou Sun, Brandon Norick, Jiawei Han, Xifeng Yan, Philip S. Yu, and Xiao Yu. Integrating meta-path selection with user-guided object clustering in heterogeneous information networks. In KDD, pages 1348–1356, 2012.

[Wang et al., 2015a] Chenguang Wang, Yangqiu Song, Ahmed El-Kishky, Dan Roth, Ming Zhang, and Jiawei Han. Incorporating world knowledge to document clustering via heterogeneous information networks. In KDD, pages 1215–1224, 2015.

[Wang et al., 2015b] Chenguang Wang, Yangqiu Song, Hao-ran Li, Ming Zhang, and Jiawei Han. Knowsim: A document similarity measure on structured heterogeneous information networks. In ICDM, pages 1015–1020, 2015.

[Wang et al., 2015c] Chenguang Wang, Yangqiu Song, Dan Roth, Chi Wang, Jiawei Han, Heng Ji, and Ming Zhang. Constrained information-theoretic tripartite graph clustering to identify semantically similar relations. In IJCAI, pages 3882–3889, 2015.

[Wang et al., 2016a] Chenguang Wang, Yangqiu Song, Hao-ran Li, Ming Zhang, and Jiawei Han. Text classification with heterogeneous information network kernels. In AAAI, pages 2130–2136, 2016.

[Wang et al., 2016b] Chenguang Wang, Yangqiu Song, Dan Roth, Ming Zhang, and Jiawei Han. World knowledge as indirect supervision for document clustering. ACM Transactions on Knowledge Discovery from Data (TKDD), 2016.

[Wang et al., 2016c] Chenguang Wang, Yizhou Sun, Yanglei Song, Jiawei Han, Yangqiu Song, Lidan Wang, and Ming Zhang. Relsim: Relation similarity search in schema-rich heterogeneous information networks. In SDM, 2016.

