論文 / 著書情報
Article / Book Information

題目(和文)	
Title(English)	Ontology-assisted Methods for the Detection and Clustering of Hierarchical Topics on the Social Web
著者(和文)	SLABBEKOORNKristian
Author(English)	Kristian Slabbekoorn
出典(和文)	学位:博士(学術), 学位授与機関:東京工業大学, 報告番号:甲第10256号, 授与年月日:2016年3月26日, 学位の種別:課程博士, 審査員:徳田 雄洋,佐伯 元司,徳永 健伸,権藤 克彦,西崎 真也
Citation(English)	Degree:Doctor (Academic), Conferring organization: Tokyo Institute of Technology, Report number:甲第10256号, Conferred date:2016/3/26, Degree Type:Course doctor, Examiner:,,,,
学位種別(和文)	博士論文
Category(English)	Doctoral Thesis
種別(和文)	論文要旨
Type(English)	Summary

（博士課程）
Doctoral Program

論 文 要 旨

THESIS SUMMARY

専攻： Department of	Computer Science 専攻	申請学位（専攻分野）： Academic Degree Requested	博士 (Philosophy) Doctor of
学生氏名： Student's Name	Kristian Slabbekoorn	指導教員（主）： Academic Advisor(main)	Takehiro Tokuda
		指導教員（副）： Academic Advisor(sub)	

要旨（英文 800 語程度）
Thesis Summary (approx.800 English Words)

With the increasing popularity of social media, efforts have been put into structuring user generated content. Traditionally, this is done using topic modeling, document clustering and community detection approaches. In this thesis we propose two methods that address weaknesses of each, by expanding text to an external ontology to reduce sparsity, introducing the concept of topic scope to prevent us from having to determine target topics, and proposing a community detection method for clustering users into labeled topic clusters. First, we give brief introductions to some natural language processing notions, and a short review of topic modeling, document clustering and community detection. We provide a survey into previous research of approaches that relate to our proposed methods, touch on their shortcomings, and introduce our proposals to improve them. We introduce two common components for our proposed methods. First, we detect entities in free text on the Social Web by linking terms to DBpedia resources using an existing entity recognition tool, DBpedia Spotlight. We subsequently apply ontological expansion to map users to classes, by collecting the full hierarchical trees of DBpedia, YAGO and Schema.org classes assigned to each resource of the DBpedia ontology.

For our first proposed method, we demonstrate the benefits of applying ontological analysis to user recommendation on Twitter. We create an initial ranking of potentially valuable users given a search keyword by analyzing user relations. For each of these users, we obtain two class taxonomies: one from the posts in which the keyword was mentioned, and one from their remaining timeline. We apply mismatch removal and generic class filtering heuristics to prune the class taxonomies. We then propose the "taxonomical similarity", a normalized measure for class overlap, which we use to determine topic consistency. In case topics are not consistent across the user's timeline, we exclude them from the recommendation ranking. In our second method, we generalize the first method to allow ontological comparison among arbitrary users. We propose a tf-idf-style weighting scheme for classes, cf-iuf, which includes a controllable "topic scope" parameter to represent hierarchical topics. Using this weighting scheme, we transform user classes into "trait vectors", which express a topic of interest at a chosen topic scope (level of generality). By calculating the cosine similarity between pairs of users, we obtain their "scoped topical similarity" (STS). We develop a community detection algorithm for weighted graphs constructed from the STS between users to cluster based on the topic scope rather than the number of topics. STS-clustering can be used in situations where the number of topics is not known beforehand. The approach additionally generates human- and machine-readable labels for clusters.

The second method is implemented into a Web application that demonstrates the real-world applicability of our method. Our system gathers users from Twitter, applies the steps of the approach to calculate STS between users, then visualizes construction and topic clustering of a graph in real-time. We analyze the algorithms used and their complexity, showing that STS-clustering has worst-case complexity O(n^3 log n) and is theoretically faster than popular k-means and LDA implementations up to dataset sizes on the order of 10^4. Benchmarks show STS-clustering takes 70.6 seconds to cluster 175 Twitter users, which is 4.7s faster than k-means, and 16.5s faster than LDA.

We evaluate the methods against manually composed ground truths. For user recommendation, we compare mainly to a follow relation-based user ranking for different keywords: applying a topic consistency check to the ranking of users allows us to obtain 7 to 17% more accurate recommendations in terms of a top-20 nDCG ranking, depending on the keyword. For topic clustering, we compare performance on Twitter user and newsgroup document ground truths to k-means and LDA. After applying a hyperparameter optimization procedure to tune our STS-clustering method, the results show an accuracy improvement of up to 26.7% over the baselines on Twitter user data. For formal newsgroup documents, the lack of context information when there are too many short posts causes our method to have poor accuracy compared to LDA. Restricting to longer documents, accuracy is again better or equivalent. For all data we can detect the correct topic content for different topic numbers or scopes (k-means fails). We do not have to pre-define the number of topics (traditional methods require this). Correct disambiguation of "Football" (Soccer) and (American) "Football" shows benefits of ontological expansion (LDA fails). We successfully derive human- and machine-readable topic labels (LDA generates loosely connected terms that machines cannot reason about).

As future work, we would like to increase the scalability of our topic clustering method by changing the community detection algorithm, and to develop methods for overlapping community detection and multi-lingual topic clustering.