Построение иерархических тематических моделей крупных конференций

Кузьмин Арсентий Александрович,
Адуенко Александр Александрович

Аспирант, аспирант

Московский Физико-Технический Институт, Москва, Россия

E-mail: arsentii.kuzmin@gmail.com

Программный комитет крупной конференции ежегодно сталкивается с задачей построения ее тематической модели. Экспертам необходимо определить положение каждого нового доклада в иерархической структуре тем конференции. Предполагая, что структура конференции меняется из года в год незначительно, предлагается построить экспертную систему для поиска наиболее подходящих тем для нового доклада с помощью тематических моделей конференций прошлых лет и методов текстового анализа.

Иерархическую структуру конференции можно представить как дерево, листами которого являются доклады, а узлами – кластеры докладов (например, темы, направления, сессии). Для документов прошлых лет известна их кластеризация на всех уровнях иерархии, поэтому классификацию новых, еще не размеченных докладов, можно рассматривать как задачу частичного обучения. Для решения подобных задач можно использовать дивизимные методы, в которых алгоритм в каждом узле иерархии выбирает наиболее подходящий дочерний кластер для нового документа. В качестве подобных алгоритмов могут быть использованы мультиклассовые SVM [1], наивные байесовские классификаторы, методы ближайших соседей [2] или взвешенные функции сходства [3].

Однако данные подходы являются жадными, и выбор кластера на верхнем уровне иерархии автоматически делает невозможным попадание доклада в кластер нижнего уровня, не являющийся его дочерним. К тому же, при небольшом размере кластеров нижнего уровня иерархии, классификация не является устойчивой, так как при добавлении нового доклада в кластер, его терминологический состав может значительно измениться, что приведет к изменению сходства данного кластера с уже находящимися в нем документами.

В статье [4] показано, что использование информации о родительских кластерах при классификации, может значительно улучшить качество классификации. В данной работе предлагается иерархическая взвешенная функция сходства документа и кластера нижнего

уровня, которая учитывает сходство документа не только с данным кластером, но и со всеми его родительскими кластерами.

Также, имея большое число кластеров и небольшое число обучающих документов, результат классификации алгоритма будет часто отличаться от экспертной классификации. Поэтому целью данной работы является предложить эксперту набор наиболее подходящих кластеров для новых документов, вместо одного наилучшего. Для этого строится оператор релевантности, возвращающий ранжированный список кластеров нижнего уровня иерархии в порядке убывания их релевантности новому документом.

В данной работе рассматривается три способа построения подобного оператора: с помощью иерархического мультиклассового SVM [1], вероятностной тематической модели ARTM [5] и предлагаемой иерархической взвешенной функции сходства. Веса данной функции настраиваются по тематическим моделям конференций прошлых лет с помощью предлагаемой энтропийной модели.

Для проверки предложенной функции сходства и сравнения ее с операторами релевантности, использующими SVM и ARTM, строится тематическая модель конференции EURO 2010 по экспертным моделям конференций EURO 2012, 2013.

Литература

1. Pei-Yi Hao. SVM classification based on support vector clustering method and its application to document categorization // Expert Systems with Applications. 2007, V. 33, №3, P. 627–635.
2. Cover T. M., Hart P. E. Nearest neighbor pattern classification. // IEEE Trans. Inform. Theory, IT-13:21–27, Jan 1968.
3. Кузьмин А. А., Адуенко А. А, Стрижов В. В. Проверка адекватности тематических моделей коллекции документов. // Программная инженерия. 2013. № 4. С. 16-20.
4. McCallum A. Improving Text Classification by Shrinkage in a Hierarchy of Classes // Proceedings of the Fifteenth International Conference on Machine Learning, San Francisco, USA, 1998, P. 359–367.
5. Vorontsov K. V. Additive Regularization of Topic Models for Topic Selection and Sparse Factorization // Statistical Learning and Data Sciences. 2015, V. 9047, P. 193–202.