распространенным алгоритмом построения тематических моделей является LDA, лишенный практически всех недостатков своих предшественников. Латентное размещение Дирихле (LDA, Latent Dirichlet Allocation) – это модель, объясняющая результаты наблюдений с помощью неявных групп, что позволяет получить объяснение, почему некоторые части данных схожи. Например, если наблюдениями являются слова, собранные в тексты, утверждается, что каждый текст представляет собой смесь небольшого количества тем и что появление каждого слова связано с одной из тем документа. LDA впервые был представлен в качестве графической модели для обнаружения тем Дэвидом Блеем, Эндрю Нг и Майклом Джорданом в 2002 году. Модель LDA устраняет такие недостатки популярных ранее моделей, как склонность к переобучению, невозможность вычислить вероятность нового документа и отсутствие закономерности при генерации документов из сочетания полученных тем.

**Литература**

1. Blei D., Ng A., Jordan M. Latent Dirichlet Allocation // Journal of Machine Learning Research. – MIT Press, 2003. – No. 3. – P. 993–1002.

2. Воронцов К. В. Математические методы обучения по прецедентам (теория обучения машин) // Курс лекций ВМК МГУ и МФТИ. – 2011.

## ИСПОЛЬЗОВАНИЕ ВЕРОЯТНОСТНОГО ЛАТЕНТНТНО-СЕМАНТИЧЕСКОГО АНАЛИЗА ДЛЯ ПОСТРОЕНИЯ ВЕРОЯТНОСТНЫХ ТЕМАТИЧЕСКИХ МОДЕЛЕЙ ТЕКСТОВЫХ КОЛЛЕКЦИЙ

М.И. Селюк, Н.Н. Шинкевич, М.В. Стержанов

Одной из важных задач автоматической обработки текстов является кластеризация текстовых коллекций. Для выявления скрытых тем в текстовых коллекциях в последнее время все чаще применяются вероятностные тематические модели. Одним из наиболее часто используемых алгоритмов для построения тематических моделей является PLSA.

Вероятностный латентно-семантический анализ (индексирование) (PLSA, Probabilistic Latent Semantic Analysis) был предложен Томасом Хоффманом в 1999 году. В основе метода лежит аспектная модель (aspect model), которая связывает скрытые (латентные) переменные тем с каждой наблюдаемой переменной словом или темой. Задача состоит в выявлении латентных переменных. Таким образом, каждый документ может относится к некоторым темам с некоторой вероятностью, что является отличительной особенностью этой модели по сравнению с походами, не основанными на вероятностном моделировании.

Достоинством метода можно считать его способность выявлять зависимости между словами, когда обычные статистические методы бессильны. PLSA также может быть применен как с обучением (с предварительной тематической классификацией документов), так и без обучения, что зависит от решаемой задачи. К недостаткам модели можно отнести склонность к переобучению и неприменимость к большим наборам данных, невозможность вычислить вероятность документа, которого нет в наборе данных, а также отсутствие какой-либо закономерность при генерации документов из сочетания полученных тем.

**Литература**

1. Manning C. D. Introduction to Information Retrieval // MIT Press, 2008.

2. Hofmann Thomas. Probabilistic latent semantic indexing // In Proc. of the SIGIR99.

## ЭЛЕКТРОННЫЕ СВОЙСТВА ДЕФЕТНЫХ СТРУКТУР ФОСФОРЕНА

В.А. Скачкова, М.С. Баранова, Д.Ч. Гвоздовский

Точечные дефекты и примеси часто оказывают значительное влияние на физические свойства материала. Экспериментальное определение дефектов обычно крайне сложное и косвенное, кроме того, требует невероятных сочетаний различных методов. При исследовании дефектов, первопринципные методы показали себя как мощный теоретический подход, достаточно надежный, чтобы использоваться как предсказательный инструмент. Проведено квантово-механическое моделирование влияния моновакансии в слое черного фосфора (фосфорене), на его