Библиографический список

1. Лорд Э.Э., Маккей А.Л., Ранганатан С. Новая геометрия для новых материалов / под ред. В.Я. Шевченко, В.Е. Дмитриенко. – М.: ФИЗМАТЛИТ, 2010. – 264 с.

2. Тушкина Т.М., Павлова Н.В. Генерация лабиринта с заданными позициями входа и выходов // МАК: «Математики – Алтайскому краю» : сборник трудов всероссийской конференции по математике, Барнаул, 29 июня – 1 июля 2016 г. – Барнаул: Изд-во Алт. ун-та, 2016. – С. 112–113.

3. Тушкина Т.М., Павлова Н.В. Вычисление вероятности выхода из лабиринта с заданными начальными и конечными точками // МАК: «Математики – Алтайскому краю» : сборник трудов всероссийской конференции по математике, Барнаул, 29 июня – 1 июля 2017 г. – Барнаул: Изд-во Алт. ун-та, 2017. – С. 281–282.

4. Смирнов В.В., Спиридонов Ф.Ф. Моделирование фракталов в Maple. – Бийск: Изд-во АлтГТУ, 2006. – 91 с.

УДК 519.6

Анализ проблематики тематического моделирования

О.Н. Половикова, Н.С. Бабкина, Л.Л. Смолякова
АлтГУ, г. Барнаул

Ключевые слова: автоматизированная обработка текст, тематическое моделирование, методы PLSA и LDA

Неотъемлемым звеном современного развития информационного общества являются исследования по созданию систем автоматизированной обработки текстов на естественном языке. Непрерывный рост объема информации, который нужно просмотреть, отобрать, проанализировать субъекту (человеку) заставляет использовать системы автоматической (или полуавтоматической) классификации, фильтрации, аннотирования, экспертной оценки и анализа обработки текстов. Несмотря на существенное множество зарекомендовавших себя на практике подходов и методов по обработки текстовых документов [1–3], данное направление исследования является актуальным. Актуальность поддерживается развитием требований по функционалу, новыми целями, которые ставит современная ситуация перед автоматизированными системами обработки текстов. А также в силу неразрешенности ряда вопросов и задач по данной проблематике.

Одними из решаемых задач в области автоматизированной обработки текстов на естественном языке (ЕЯ) является задачи, связанные с выделением тематик для текстовых документов в коллекции. И, в частности, перспективная и малоизученная задача построения тематических деревьев. Тематическое моделирование (и в частности, иерархическое тематическое моделирование) выросло на исследованиях по классификации текстовых ресурсов. Поэтому базовые идеи реализации методов и алгоритмов классификации текстов заложили методологическую основу инструментария для определения тематик коллекций документов. Также актуальными остались недостатки и проблемы реализации применяемых методов и их адаптации к конкретным практическим задачам.

Основными современными методами тематического моделирования являются [4, 5]:

1) вероятностный латентный семантический анализ (Probabilistic Latent Semantic Analysis, PLSA), который реализуется итерационным EM-алгоритмом;

2) латентное размещение Дирихле(Latent Dirichlet allocation, LDA) является развитием метода PLSA. Данный подход учитывает дополнительные предположения относительно строящейся модели, что позволяет не перестраивать заново всю модель при добавлении нового документа в коллекцию, а также избежать переобучения модели (недостатки PLSA).

В данном исследовании приведен анализ некоторых основных проблем методов автоматизированного построения тематик текстов, а также способы или возможные направления их решения. Проблемную область можно условно разделить на три направления:

1. Технические проблемы, связанные с ресурсными ограничениями.

Прямая реализация алгоритмов (метода PLSA и LDA) может привести к высокой размерности пространства параметров. Данная проблематика изучена и решается методами машинного обучения путем отбора значимых признаков для пространства параметров, формированием новых признаков путем трансформации первоначальных, а также использованием регуляторов. Проблема сложных и ресурсоемких затрат вероятностных моделей на каждом итерационном шаге может частично разрешаться методами сэмплирования. Сэмплирование является одним из основных инструментов приближённого вывода в сложных вычислительных моделях.

2. Лингвистическая специфика интерпретаций терминов ЕЯ:

– неоднозначность интерпретации слов ЕЯ (синонимы, антонимы);

– изменение слов по грамматическим категориям (разные окончания), присутствие в тексте неинформативных единиц (предлогов, союзов и т.д.),

– присутствие в тексте слов, которые отвечают за индивидуальность документа (шум) и за особенность коллекции (фон), а не за его тематику [6].

Данный класс проблем решается на этапе предобработки текстовой коллекции. Основным методам предобработки является нормализация данных (лемматизация, стёмминг, отбрасывание *стоп-слов*). Для замены антонимов и синонимов используются специальные словари по предметной области. Учет специальных слов, которые не относятся к тематике документов, может быть реализован на основе робастных тематических моделей, в которые добавляется шумовая и фоновая составляющие.

3. Вопросы обоснования применимости того или иного метода или алгоритма, а также интерпретация результатов их работы на коллекциях.

Вопросы обоснования метода и интерпретации его результатов остаются открытыми. Частично задача повышения интерпретируемости тематической модели может быть решена путем подбора специальных регуляторов в качестве ограничителей на множество допустимых решений. А также используя комбинированных критериев проверки достоверности построенной модели [5].

Проведенное исследование выявило проблемную область тематического моделирования, несмотря на весомое количество работ в данном направлении и огромное количество построенных моделей.

Библиографический список

1. Автоматическая обработка естественного языка: учебное пособие / Луканин А.В. – Челябинск: Издательский центр ЮУрГУ. – 2011.

2. Автоматическая обработка текстов на естественном языке и компьютерная лингвистика : учеб. пособие / Большакова Е.И., Клышинский Э.С., Ландэ Д.В., Носков А.А., Пескова О.В., Ягунова Е.В. – М.: МИЭМ. – 2011.

3. Лушев А.А., Половикова О.Н. Формальная и контекстная проверка текстовых документов // Труды семинара по геометрии и математическому моделированию. – 2016. – №2. – С. 36–38.

4. Воронцов К. В. Вероятностное тематическое моделирование [Электронный ресурс]. URL: www.machinelearning.ru/wiki/images /2/22/Voron-2013-ptm.pdf.

5. Воронцов К. В., Потапенко А. А. Многокритериальная регуляризация вероятностных тематических моделей для улучшения интерпретируемости тем и определения числа тем // Диалог–2014.

6. Chemudugunta C., Smyth P., Steyvers M. Modeling general and specific aspects of documents with a probabilistic topic model // Advances in Neural Information Processing Systems. — MIT Press, 2006. – Vol. 19. – P. 241–248.

УДК 004.738.5

Возможности логического языка для решения задач на основе программной генерации фактов

О.Н. Половикова
АлтГУ, г. Барнаул

Возможности логического языка для решения задач на основе генерации процедуры правил

Логические языки нашли широкое применение во многих прикладных областях, в том числе и для решения специфических задач, где требуется применять узко настраиваемые модели знаний и аппарат получения новых знаний. Среди логических языков особое место занимает язык Пролог (Акторный пролог, Visual Prolog и др. среды), как широко используемый инструмент декларативного логического программирования. Такая применимость обусловлена поддержкой, помимо логического программирования, объектно-ориентированной парадигмы, программирования, управляемое шаблонами.

Программирования на языке логики, во-первых, привлекает к процессу разработки не только программистов и инженеров знаний, но и других специалистов с различными научными интересами, например, математиков. Во-вторых, определяет одно из перспективных направлений использования Пролог-систем – в качестве аппарата для решения логических задач. **В работе анализируется один из подходов построения новых элементов базы знаний – генерацию состояний.** Динамическое формирование множества возможных состояний объекта и способов его поведения позволяет выполнить поиск решения задачи в тех случаях, когда либо все дерево состояний хранить нецелесообразно исходя из специфики решения, либо предварительное построение и хранение такого дерева может привести к комбинаторному взрыву.

Кроме этого этап динамической генерации состояний будет ключевым в решении тех задачах, где пользователь программы управляет