

Добрый день, уважаемые члены комиссии!

Мой проект связан с предиктивной аналитикой перспективных технологий на основе анализа патентов. Анализируя распределение патентов по классам патентного классификатора, предлагается качественно оценить вероятность структурных изменений в классификаторе, выделить точки его дальнейшего развития, а следовательно, прогнозировать направления развития технологий во времени.

Патентные документы являются отражением технологического развития нашего мира. А в силу того, что в последнее время количество зарегистрированных патентов растет экспоненциально, в сфере анализа и обработки больших данных такой объем насыщенной информации является очень ценным ресурсом. Мало того, что возникает все большая потребность классифицировать увеличивающийся объем патентов, так и имеющийся классификатор может со временем устаревать и не отражать действительности в развитии технологий. При этом также важно превентивно выявлять изменения, которые происходят в сфере технологического развития на ранних этапах, чтобы учесть все риски и новые возможности в областях, связанными с исследуемыми технологиями. Таким образом предлагаемая программная система решает сразу несколько задач.

Лист 1. Схема структурная. На листе 1 представлена структурная схема предлагаемой системы. Она включает в себя модули построения моделей, пользовательского интерфейса, модуля визуализации, модуля генерации отчета и модуля хранения данных. В представленной работе решаются следующие задачи:

- Анализ существующих методов прогнозирования, включая методы векторизации и кластеризации.
- Выбор методов и моделей обработки данных для использования в системе
- Разработка алгоритма формирования.
- Разработка структуры программного обеспечения и ее реализация.
- Разработка технологии тестирования и комплексное тестирование системы

Лист 2. Диаграмма вариантов использования. На листе 2 представлена диаграмма вариантов использования системы. Она отражает варианты взаимодействия пользователя и включает настройку системы, загрузку данных, формирование наборов данных, построение моделей патентного ландшафта и построение прогнозов.

Лист 3. Анализ подходов к извлечению информации. Для того чтобы определить, какую технологию использовать для решения поставленной задачи, выполнен анализ подходов к прогнозированию появления новых технологий с учетом специфики патентных документов. Были рассмотрены подходы на основе кластеризации, подходы на основе ресурсных сетей, а также подходы на основе машинного обучения. На листе 3 приведены обобщающие таблицы с итогами анализа. Так как каждый из подходов имеет свои

преимущества и недостатки, было решено выбрать гибридный метод, который максимизирует преимущества каждого.

Лист 4. Общая схема подхода. На листе 4 представлена предлагаемая концепция. Загруженные патентные документы формализуются и разбиваются по семантическим структурам: данные о патенте, о его классах в патентном классификаторе и данные о цитировании. Далее из всего множества патентных документов создаются наборы данных, которые удобны для того, чтобы оперировать группами патентов, выбранных фильтрами, например по классам и по времени их появления. Следующим шагом на определенном наборе данных строится модель патентного ландшафта через кластеризацию с оптимизацией параметров кластеризации для того, чтобы максимально точно отразить структуру патентов в выбранном наборе данных. После этого с известными параметрами кластеризации создается уже прогнозная модель на наборе данных из следующего промежутка времени. Наконец две модели сравниваются, количественно оцениваются и создается прогноз.

Лист 5. Концептуальная модель предметной области. На базе предлагаемого подхода построена концептуальная модель предметной области, изображенная на листе 5. Модель отражает основные сущности, которыми оперирует система. К примеру, это патент, который получен из документа, содержит цитирования, имеет патентные классы и используется менеджером векторизации для получения его векторного представления.

Лист 6. Схема алгоритма прогнозирования. На листе 6 представлены главные алгоритмы прогнозирования, которые соответствуют подходу на листе 3. Здесь представлен алгоритм формирования модели патентного ландшафта, который использует векторные представления патентов, полученных с помощью SentenceBert для кластеризации методом DBSCAN с оптимизацией параметров. Алгоритм построения прогнозной модели, который похож на первый алгоритм, но использует уже известные параметры и алгоритм генерации прогнозов, который на основе сравнения двух моделей создает гипотезы и их вероятности.

Лист 7. Схемы алгоритмов модулей. На листе 7 представлены схемы алгоритмов основных модулей системы, такие как алгоритм формирования матрицы ассоциаций, которая используется для генерации гипотез, алгоритм парсинга документов, алгоритм ассоциации кластеров с классами, который используется при построении моделей, а также алгоритм генерации отображения патентного ландшафта методом UMAP и выделения ключевых слов методом LDA.

Лист 8. Диаграмма классов предметной области. Предлагаемый подход реализован в на языке Python, и на листе 8 представлены классы сущностей, разделенные на пакеты.

Лист 9. Алгоритм тестирования. Был также разработан алгоритм тестирования модели прогнозирования и оценки качества, который представлен на листе 9. Он включает в себя комплексную оценку работы системы, а также оценку результатов прогнозирования.

Лист 10. Результаты тестирования. Выполнено тестирование системы, подтвердившее корректную работу программы. Помимо тестирования проведены эксперименты по созданию прогнозов для некоторого набора патентных документов. Все результаты представлены на листе 10, на котором также отображены формы интерфейса разработанной системы. К примеру, в выбранном временном интервале можно заметить смещение фокуса патнетного класса G01L «Регулирующие и управляющие системы общего назначения» с производства на обработку в контексте информации, а также некоторое сближение к классу A61B «Диагностика; хирургия; опознание личности» что говорит о цифровизации сферы медицины и подтверждается научными статьями. Тестирование точности построения гипотез показало, что система с достаточным уровнем достоверности может являться достойным вспомогательным инструментом в сфере прогнозной патентной аналитики. Результаты работы были включены в состав прототипа интеллектуальной унифицированной системы поддержки принятия стратегических решений в рамках реализации программы развития «Приоритет 2030».

Следующее видео показывает пример полного цикла создания прогноза.

Спасибо за внимание, доклад окончен.