1 слайд

Доброе утро уважаемые члены комиссии. Вашему вниманию представляется выпускная квалификационная работа бакалавра на тему “рекомендательная система новостей на основе нечеткой кластризации”.

2 слайд

За последние несколько десятилетий, с появлением Youtube, Amazon, Netflix и многих других подобных веб–сервисов, системы рекомендаций стали занимать все больше места в нашей жизни. Начиная с электронной и заканчивая рекламой в Интернете.

Многие современные сервисы создают рекомендательные системы, которые основываясь на информации о пользователе и его поведении в системе, пытаются определить какие объекты ему интересны, будь то товары, новости, услуги  и т.д. Яркими примерами служат такие сервисы или сайты, как «КиноПоиск», «Яндекс.Дзен», «Яндекс.Новости» и многие другие. К примеру «КиноПоиск» --- российский веб-сайт, предлагающий пользователю к просмотру фильмы на основе его предпочтений, а «Яндекс.Дзен» --- веб-сайт и расширение для браузера от компании «Яндекс», ищущее в интернете информацию, которая может быть интересна пользователю, и собирающее ее в персональную ленту.

3 слайд

Цель работы — разработать и реализовать рекомендательную систему новостей на основе нечеткой кластеризации.

Для достижения поставленной цели необходимо решить следующие задачи:

- провести анализ предметной области;
- проанализировать подходы к реализации рекомендательных систем ;
- в результате полученных во время анализа данных разработать рекомендательную системы на основе нечеткой кластеризации;
- реализовать алгоритм нечеткой кластеризации;
- реализовать рекомендательную систему в программном продукте;
- провести исследование работоспособности реализованной рекомендательной системы и алгоритма нечеткой кластеризации.

4 слайд

Для достижения результата в виде рекомендаций новостей потребуется разработать рекомендательную систему на основе нечеткой кластеризации, на вход которой будут поступать новости, к которым алгоритм должен предложить рекомендации. Подобная работа отражена на схеме, приведенной на экране.

5 слайд

Учитывая то, что рекомендательная система разрабатывается для новостного сайта, можно выделить требования, что система должна работать зачастую с новостями, которые только были опубликованы и не имеют просмотров, также требуется учитывать, что пользователь имеет свой уникальный набор интересов и тем для просмотра новостей. Исходя из данных требований, был выбран метод фильтрации на основе содержания. Он позволяет предлагать объекты пользователю, не основываясь на интересах других людей.

6 слайд

Взглянув на таблицу можно сделать вывод о том, что выбор модели Гауссовой смеси в качестве алгоритма нечеткой кластеризации для рекомендательной системы является наилучшим решением, поскольку данный метод показывает лучший результат на большом количестве данных в плане качества работы. Скорость работы компенсируется качеством, к тому же модель требуется обучить на большом наборе данных лишь один раз перед запуском системы.

7 слайд

На данном слайде можно увидеть декомпозицию разработанной рекомендательной системы.

Как видно разработанная система состоит из следующих этапов

1. предобработка входных данных
2. векторизация предобработанных данных
3. понижение размерности векторизованных данных
4. разделение новостей на кластеры с использованием алгоритма нечеткой кластеризации
5. реомендация новостей на основе данных, выделенных в кластеры.

В последующих слайдах будет подробно описан каждый этап приведенной диаграммы.

8 слайд

Данный этап предназначен для подготовки к дальнейшему обучению входных данных. На вход поступают структуры из json-файла, которые содержат информацию о новостях.

Для того, чтобы корректно произвести векторизацию и последующее обучение модели, данные следует предобработать следующим образом:

• объединить столбцы с заголовком, абстрактным описанием и самой новостью;

• удалить все символы, кроме кириллических;

• удалить все ”стоп-слова”;

• провести лемматизацию.

Лемматизация это процесс, использующий лексикон и морфологический анализ слов, который позволяет приводить слова к их словарной форме.

Все выполняемые на данном шаге действия и лемматизация в том числе позволяют улучшить последующее качество модели нечеткой кластеризации.

9 слайд

На данном этапе происходит векторизация ранее предобработанных данных. Векторизация --- это процесс при котором предобработанная новость становится вектором, где каждое значение вектора это слово, преобразованное с использованием терм документной частоты формула которой приведена на экране.

10 слайд

На данном этапе происходит понижение размерности матрицы полученной на предыдущем этапе, поскольку при векторизации была создана матрица признаков с большим количеством столбцов, большинство из которых не несут полезной информации. Если говорить просто, то понижение матрицы позволяет извлечь всю полезную информацию, значительно скоратив общее количество информации. Если сложно то сингулярное разложение работает по следующему

принципу любая матрица, в нашем лучае векторизованная матрица предобработанных данных представляется в виде произведения трех матриц:

$$\mathbf{X=U\Sigma V^*},$$

где U — унитарная матрица порядка m; Σ — матрица размера m x n, на главной диагонали которой лежат неотрицательные числа,
называющиеся сингулярными (элементы вне главной диагонали равны нулю — такие матрицы иногда называют прямоугольными диагональными матрицами); $V^*$ — эрмитово-сопряжённая к V матрица порядка n. m столбцов матрицы U и n столбцов матрицы V называются соответственно левыми и правыми сингулярными векторами матрицы X. Для задачи редукции количества измерений именно матрица Σ, элементы которой, будучи возведенными во вторую степень, можно интерпретировать как дисперсию, которую «вкладывает» в общее дело каждая компонента, причем в убывающем порядке: $\sigma_1 \geq \sigma_2 \geq \ldots \geq \sigma_{noise}$. Поэтому при выборе количества компонент при SVD ориентируются именно на сумму дисперсий, которую дают учитываемые компоненты, но об этом будет рассказано в исследовательском разделе.

11 слайд

После понижения размерности выполняется этап нечеткой кластеризации новостей из полученной матрицы. В качестве алгоритма нечеткой кластеризации используется модель Гауссовой смеси (Gaussian Mixture Model (GMM)). В данном алгоритме каждый кластер представляется параметрическим распределением, а весь набор данных моделируется смесью этих распределений. Нахождение данных распределений осуществляется методом максимизации ожидания. Исходя из названия становится ясно, что алгоритм состоит из двух частей, а именно вычисления ожидания (E), которое приведено в формулах 17, и вычисления максимизации (M).

По схеме алгоритма шага ожидания видно, что мы получаем матрицу в которой строки — это каждый элемент данных, а столбец представляет кластер, следовательно каждый элемент данной матрицы это вероятность принадлежности элемента данных к столбцу. После того как алгоритм сойдется, данные из этой матрицы будут использованы в качестве предсказания точки кластера.

А на шаге максимизации происходит вычисление нового вектора средних для каждого столбца, в обновление матрицы ковариаций для каждого столбца и обновление список отношений количества элементов в кластере ко всем элементам данных.

Сам же алгоритм нечеткой кластеризации это итеративный процесс повторного выполнения этих 2 шагов, и последующее присвоение данным кластеров с наибольшей вероятностью принадлежности.

12 слайд

Заключительным этапом разработки приложения, является создание рекомендательной системы новостей на основе данных, полученных после проведения нечеткой кластеризации, а именно предоставление пользователю набора новостей из того же кластера, что и выбранная им новость и кластеров, центры которых наиболее близки к текущему. Результатом данного этапа является рекомендательная система, предоставляющая пользователю рекомендации новостей на основе его персональных предпочтений.

15 слайд

Как можно заметить на графике при количестве компонент равным 6700 доля объясненной дисперсии превышает 0.95, что является хорошим показателем и позволяет существенно сократить затраты по памяти. По своей сути долю объясненной дисперсии можно сравнить с примеров, когда человек приезжает в другой город и хочет узнать весь город. В данном случае, чтобы понять почти весь город не требуется знать все его улицы, достаточно обладать

информацией лишь о самых главных из них, именно это и отражает доля объясненной дисперсии другими слвоами, количество информации сохранившееся после сокращение общего количества компонент.

13 слайд

Поскольку у нас нет априорных данных о принадлежности новостей к категориям, требуется выбрать такой критерий оценки, который позволит оценить насколько объект похож на свой кластер по сравнению с другими кластерами. Данную задачу наилучшим образом решает метод оценки качества под названием Силуэт (Silhouette). Метод силуэтов — способ изучения разделительного расстояния между результирующими кластерами наблюдений, данная мера имеет диапазон [-1, 1]. Коэффициенты силуэта около +1 указывают на то, что образец находится далеко от соседних кластеров. Значение, близкое к нулю указывает, что выборка находится на границе принятия решения между двумя соседними кластерами или очень близко к ней, а отрицательные значения указывают на то, что эти выборки могли быть назначены неправильному кластеру.

Для достижения наилучшего результата нечеткой кластеризации требуется варьировать ключевой параметр, которым является количество кластеров, требуется получить такое количество кластеров при котором оценка методом Силуэта будет давать наивысший результат (быть как можно более близким к 1). Для достижения необходимого результата был проведен эксперимент при котором варьировалась количество кластеров.

Как можно заметить на графике наилучшая оценка достигается при количестве кластеров равным 380, данное количество кластеров и было выбрано в качестве парамертра для обучения модели.