Практическая работа № 3

«Модельные методы рекомендательных систем на основе сингулярного разложения и кластерного анализа»

по дисциплине «Разработка обеспечивающих подсистем систем поддержки принятия решений»

Цели: приобрести навыки реализации модельных коллаборативных методов рекомендательных систем, основанных на поиске скрытых факторов методами сингулярного разложения и кластерного анализа.

Задачи:

1) создать программную реализацию модельного метода рекомендательной системы (РС), основанную на сингулярном разложении и кластерном анализе и включающую:

– актуальную предметную область для применения РС (вроде маркетплейса, медиа ресурсов, соц. сетей, экономической сферы и т.д.) и набор начальных данных для неё (опрос покупателей, статистка за временной период и т.д., см. Примечание 1);

– привести входные данные к матричному виду для выполнения её сингулярного разложения вида $A = UDV^T$ и получения на его основе предпочтений пользователей, характеристики объектов и сингулярных значений как значимость воздействия факторов на пользователей, интерпретировать эти факторы для человеческого восприятия (см. Примечание 2);

– произвести измерение качества рекомендаций на основе метрик качества (вроде MSE, MAE и т.д.) и реальных оценок полученных пользователями рекомендаций;

– выполнить кластеризацию пользователей или их рекомендаций (результатов предыдущих пунктов задания) с помощью EM-алгоиртма или иерархической кластеризации;

2) в качестве дополнительного задания (на доп. баллы) выполнить:

– реализации алгоритма Funk SVD с минимизаций найденной ошибки (см. Примечание 3) для работа с явными данными пользователей и алгоритма SVD++ с проведением регуляризации для компенсирования переобучения для работа с неявными данными;

– выполнить оба алгоритма кластерного анализа, приведённых выше, и реализовать для них несколько методов измерения расстояния между кластерами (см. Примечание 4), сравнить результаты для разных методов на одних и тех же классах (в данном контексте лучше сказать классах, так как кластерах, т.к. последних может быть нефиксированное количество), а также для одного метода на разных кластерах, результаты привести в виде графиков и описательного вывода.

ПРИМЕЧАНИЕ:

1. Для выполнения сингулярного разложения необходимо иметь входные данные в матричном виде (так как раскладывается именно матрица), поэтому для работы наиболее подойдут данные в форматы простой таблицы, датафрейма или матрицы, полученной от другого приложения или модуля, если же данные в другом формате (векторном, графическом или текстовом), то их необходимо преобразовать к матричному виду.

2. Сингулярные значения матриц, полученные в результате сингулярного разложения входных (анализируемых) данных, как правило, несут в себе определённый (скрытый) смысл (но могут не иметь его вообще, поэтому нужно ещё и анализировать значения полученных данных и их положение относительно друг друга), который необходимо интерпретировать понятным для человека образом благодаря знаниям предметной области (или областей), которым принадлежат эти данные, для этого необходимо проанализировать предметную область выполняемой задачи, куда входят базовые постулаты (определения, аксиомы, теоремы, правила и т.п.), выводы, взаимосвязи с другими областями и прикладное применение.

3. Минимизацию ошибки необходимо проводить до тех пор, пока пользы от неё больше, чем вреда:

a) проверять результаты минимизации на каждом шаге итерации нужно на всех значениях матрицы, которые она затрагивает;

b) нужно отслеживать историю изменений значений, чтобы ошибка снова не начала возрастать;

c) если минимизация имеет слишком малую тенденцию или изменения вообще прекратились, то её необходимо завершать, признаком может являться изменение в сотых и иногда десятых частях после наименьшего разряда в числе (например, для числа 3,14 числа 3,1415 и 3,141 уже могут быть не ликвидными, исходя из задачи, поэтому смысла их высчитывать может и не быть), для более доказательного предела вычислений можно воспользоваться теорией погрешности вычислений (вычислительная математика).

4. Разные методы поиска кластеров (метод ближнего соседа, наиболее удалённых соседей, невзвешенный и взвешенный центроидный метод и т.д.) могут совершенно по-разному определять кластеры (искать центроиды, размер кластеров, локализацию и в том числе количество этих кластеров), задачей аналитика данных в таком случае является опытный подбор методов кластеризации исходя из условий задачи и особенностей работы методов (стремление поиска наибольшего количества кластеров, наименьшего, разреженности кластеров или их пересечений), поэтому для поиска более подходящего решения задачи необходимо сравнивать разные методы на одних и тех же данных, чтобы подобрать наиболее актуальные методы, в некоторых случаях результаты методов имеет смысл совмещать.