Разработка математической модели и алгоритмов для контентных и гибридных Рекомендательных Систем

Понизовкин Д. М.

ИПС им. А. К. Айламазяна РАН

24 октября 2012 г.

Типы рассматриваемых систем

- пользователь системы, имеющий сови вкусы и предпочтения
- объект системы: книга, фильм, новостная лента и т.п.
- терм — информационная единица, например слово, жанр, название стиля и т.п.
- характеристика пользователя — его оценка на объект или взвешенный терм
- информация о вкусах пользователя представляется в виде векторов характеристик
- информация об объекте представляется в виде векторов характеристик

Словесное описание проблемы

- новое поколение цифровой техники — новые возможности, новые проблемы;
- проблема *поиска* — необходимо уметь находить среди огромного числа данных информационные объекты, удовлетворяющие вкусам пользователя
- проблема *прогнозирования оценки пользователя* — необходимо уметь определять оценку пользователя на интересующий его объект
- проблема *кластеризации* — необходимо уметь объединять/разделять по группам(кластерам) информационные объекты

Области, в которых встречаются подобные задачи

- Информационный Поиск — дисциплина, занимающаяся поиском неструктурированных данных(в основном в текстовом представлении), соответствующих установленному запросу
- Data Mining —мультидисциплинарная область, возникшая и развивающаяся на базе таких наук как прикладная статистика, распознавание образов, искусственный интеллект, теория баз данных и др.
- Рекомендательные Сервисы:
  - коллаборативная фильтрация — методология прогнозиварония оценки пользователя на объект, основанная на информации о предпочтениях пользователей
  - контентная методика — рекомендации производятся на основе информации о «контенте» (содержании) объекта
- Сервисы по поддержке принятия решений жюри — составление объективной оценки

Актуальность работы

- Необходимость создать обоснованные методы по оценке и решению задач
- Возможность создания новых концепций для хранилищ данных со встроенной рекомендатеьной системой
- Спрос пользователей на существование хорошего рекомендательного сервиса
- Растущее число публикаций по теме

Контентный метод для решения задачи рекомендации

- Объект предствляется вектором взвешенных термов:
  - частота терма tf
  - инверсированная частота терма itf
  - $tf - idf_t = tf_t \times idf_t$
- M — множество векторов-контентов объектов, зарегестрированных в системе
- N — множество векторов-контентов объектов, ранее выбранных пользователем(высоко оцененных)
- решение — n ближайших объектов из множества $M \setminus N$ к объектам множества N
- в качестве расстояния между объектами o_1, o_2 используется

$$cos(o_1, o_2) = \frac{\sum_{i}^{K}(w_i^{o_1} \times w_i^{o_2})}{\sqrt{\sum_{i}^{K}(w_i^{o_1})^2} \times \sqrt{\sum_{i}^{K}(w_i^{o_2})^2}} \quad (1)$$

Коллаборативный метод решения задачи прогнозирования оценки

- активный пользователь u_a — пользователь, чья оценка прогнозируется двумя пользователями:
- функция схожести sim — функция, вычисляющая меру схожести между
  - коэффициенты корелляции
  - косинус угла между векторами оценок
  - Евклидово расстояние
- сосед — пользователь u такой, что $sim(u_a, u) < threshold$
- прогнозная оценка

$$p = \sum_{u \in N} \frac{r_u \times w_u}{|N|} \tag{2}$$

  - N — множество соседей
  - r_u — оценка пользователя на прогнозируемый объект
  - w_u — вес пользователя

Примеры РС

- Amazon — сервис, производящий рекомендацию на книги, которые понравятся пользователю, основываясь на информации о книгах и о том, которые он предпочиатет;
- IMDB — сервис, производящий рекомендацию на фильмы, которые понравятся пользователю, основываясь на информации о фильмах и о том, которые он предпочиатет;
- LastFm — сервис, производящий рекомендацию на музыкальные произведения, которые понравятся пользователю, основываясь на информации о музыкальных произведениях и о том, которые он предпочиатет;
- MovieLens — сервис, производящий рекомендацию на фильмы, которые понравятся пользователю, исходя из оценок на знакомые ему фильмы и информации об оценках других пользователей;
- IMHO — мультикультурный рекомендательный сервис, основанный на методах коллаборативной фильтрации;

Способы оценки качества рекомендательных сервисов

- примеры самых распространенных критериев
  - точность

$$Pre = \frac{N_{relevant}}{N_{recomended}} \tag{3}$$

    - $N_{relevant} = |Rec \bigcap Sel|$
    - $N_{recomended} = |Rec|$
    - Rec — множество рекомендованных объектов
    - Sel — множество объектов, ранее выбранных пользователем

  -

$$MAE = \frac{\sum\limits_{(u,o)} |p_u^o - r_u^o|}{K} \tag{4}$$

    - p_u^a — прогнозная оценка пользователя на объект o
    - r_u^o — реальная оценка пользователя на объект o
    - K — общее количество выставленных оценок всеми пользовтелями

Проблемы существующих алгоритмов

- алгоритмы основаны на неявных предположениях
- отсутствует анализ предельных возможностей PC
- выбор критериев произвольный
  - *MAE*
    - использование в системах с балльной шкалой оценок
    - пользователь хочет знать, стоит ли ему покупать объект; реальная оценка будет 4, прогноз: 3 или 5
  - *Pre*
    - $Rec = \{e, f, g, h\}$
    - $Sel = \{a, b, c\}$
    - объекты $\{e, f, g, h\}$ понравятся пользователю и $\frac{N_{relevant}}{N_{recomended}} = 0$

Цели работы

- Обосновать критерий качества работы РС на основе метематической модели взаимодействия пользователя с РС
- Построить модели и алгоритмы для решения задач
  1. выработки рекомендаций
  2. выработки рекомендаций для дальнейшего использования в системах жюри
  3. прогнозирования объективной оценки в системе жюри

Обозначения и определения задачи 1, 2

- множество характеристик объектов C_T
- множество характеристик пользователей C_U
- контент объекта $(c_1^t, .., c_M^t)$(далее объект)
- контент пользователя $(c_1^u, .., c_N^u)$(далее пользователь)
- T — множество объектов
- U — множество пользователей
- профиль пользователя $prof(u) \subset T$
- распределение q — множество профилей всех пользователей, для которых производится рекомендация
- $char :: C_U \to C_T$ — определяется как максимальное соответствие характеристик объекта и пользователя
- $d_T :: T \times T \to \mathbb{R}$ — расстояние на множестве объектов
- $d_{UT} :: U \times T \to \mathbb{R}$ — расстояние между пользователем и объектом

Решение задачи 1

- множества U и T являются множествами различной природы
- расстояние между элементом множества $u \in U$ и $t \in T$ определим как

$$d_{UT}(u,t) = inf\{d_T(x,t) : x \in corr(U)\} \tag{5}$$

где $corr(U) = maxchar(c_u)$

Критерий решение задачи 1

качество распределения s

$$es(s) = \sum_{i=1}^{K} \frac{\sum\limits_{t \in prof(u_i)} d_{UT}(u_i, t)}{|prof(u_i)|} \to \max_t \tag{6}$$

- $K = |q|$

Алгоритм решения задачи 1

- **Шаг 0**. Инициализация алгоритма.
  - выбор случайного распределения s, установка начальной температуры $TEMP$ и коэффициента $r \in (0, 1)$.
- **Шаг 1**. Если $TEMP$ близко к нулю, то алгоритм заканчивается.
- **Шаг 2**. Расчет значения функции $ed(s) = val$;
- **Шаг 3**. Выбор соседнего распределения s';
- **Шаг 4**.
  - Если $(ed(s) > ed(s')) \Rightarrow s := s'$
  - Иначе если $(exp(\frac{val' - val}{TEMP}) < random(0, 1)) \Rightarrow s := s'$
- **Шаг 5**. $TEMP := TEMP \cdot r$ (понижение температуры).
- **Шаг 6**. переход к **Шаг 1**

Пример

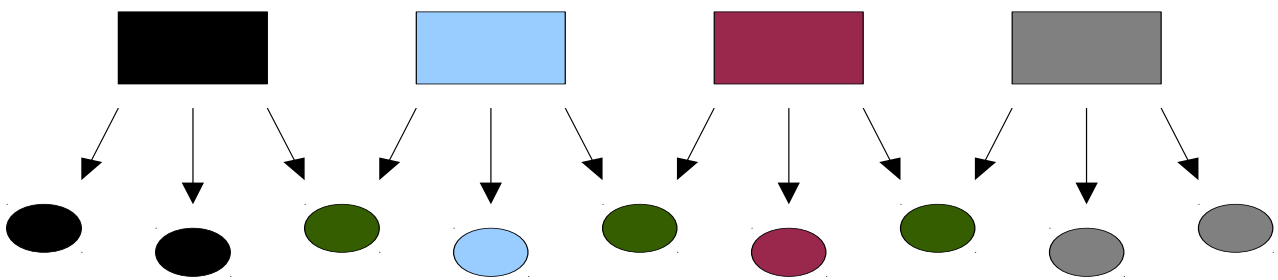
- каждый пользователь запрашивает по одному объекту
- бинарная шкала оценок для характеристик пользователей и объектов $\{0, 1\}$
- пользователи описываются двумерным вектором характеристик (c_1^u, c_2^u)
- объекты описываются двумерным вектором характеристик (c_1^t, c_2^t)
- $corr(a, b) = (a, b)$
- d_{UT} — расстояние Хэмминга при определенных условиях

Выбор решения примера по значению es

- два пользователя
  1. $u_1 = (1, 0)$
  2. $u_2 = (0, 1)$
- два объекта:
  1. $t_1 = (0, 1)$
  2. $t_2 = (1, 0)$
- четыре вараинта решения
  1. $s_1 = \{(u_1, t_1), (u_2, t_1)\}$; $es(s_1) = 1$
  2. $s_2 = \{(u_1, t_1), (u_2, t_2)\}$; $es(s_1) = 0$
  3. $s_3 = \{(u_1, t_2), (u_2, t_1)\}$; $es(s_1) = 2$
  4. $s_4 = \{(u_1, t_2), (u_2, t_2)\}$; $es(s_1) = 1$
- s_3 — решение

Решение задачи 2

- отличия от задачи 1 — требование к наличию общих объектов у каждой пары пользователей(связный граф) для нахождения корреляции между оценками пользователей



Решение задачи 2

- нужно изменить выбор начального распределения, учитывая связность графа
- нужно изменить выбор соседнего распределения, учитывая связность графа

Обозначения и определения задачи 3

- к характеристикам объекта добавим оценку пользователя
- профиль пользователя $(p_1^u, .., p_N^u)$ — вектор оценок
- U — множество пользователей
- $proj(u_1, u_2) : U \times U \to U$ — проекция профиля u_1 на u_2

Решение задачи 3

- объективный вектор оценок — вектор оценок $u_o : \sum_{i=1}^{|U|} d_U(u_o, u_i)$
- агломеративная кластеризация — объединение в один большой кластер
- составление профиля объективного пользователя:
  - **Шаг 0**
    - u_1, u_2 — пользователи, расстояние между которыми минимально
    - $u_o = proj(u_1, u_o) \bigcup proj(u_2, u_o)$ — объективный профиль
    - $(u_1, u_2) \in C$ — кластер
  - **Шаг 1** стоп, если длина профиля объективного пользователя равна количеству объектов
  - **Шаг 2** находим u, ближайшего к C
  - **Шаг 3** $u_o = proj(u, u_o)$ — проецируем профиль u на u_o
  - **Шаг 4** $u \in C$ — добавляем пользователя в кластер
  - **Шаг 5** переход к 1-ому шагу

Результаты задача 1

- сравнение с результатами статьи Cantador, Iván and Bellogín, Alejandro and Vallet, David, Content-based recommendation in social tagging systems. ACM RecSys '10, pp. 237-240, 2010
  - база lastFM
  - шесть различный функций расстония
  - три критерия
    1. точность для количества рекомендованных объектов, равного 5, 10 и 20
    2. средняя точность(MAP)
    3. линейно взвешенная релевантность (NDCG)

мера сходства	P@5	P@10	P@20	MAP	NDCG
tf_u	0.028	0.021	0.014	0.011	0.085
tf_i	0.0001	0.0001	0.000	0.0001	0.027
tf-cosine	0.234	0.109	0.059	0.041	0.202
tf-idf	0.226	0.105	0.075	0.051	0.216
bm25	0.062	0.047	0.035	0.020	0.141
bm25-cosine	0.364	0.260	0.172	0.145	0.390

мера сходства	P@5	P@10	P@20	MAP	NDCG
tf_u	0.580	0.562	0.532	0.428	0.572
tf_i	0.114	0.110	0.100	0.068	0.149
tf-cosine	0.384	0.384	0.354	0.265	0.469
tf-idf	0.254	0.232	0.208	0.170	0.270
bm25	0.142	0.138	0.126	0.032	0.164
bm25-cosine	0.270	0.258	0.240	0.184	0.294

Результаты по задаче 2, 3

- база EDU
- критерий оценки
  - идеально — количество пар (пользователь, объект), расстояние d между которыми $d = 1$
  - хорошо — количество пар (пользователь, объект), расстояние d между которыми $0.99 \geq d < 1$
  - средне — количество пар (пользователь, объект), расстояние d между которыми $0.9 \geq d < 0.99$

Результаты по задаче 2, 3

- результат системы

идеально	хорошо	средне
34	36	15

значение функции критерия 0.931

- полученный результат

идеально	хорошо	средне
34	46	5

значение функции критерия 0.962

Готовые публикации

- Понизовкин Д. Построение оптимального графа связей в системах коллаборативной фильтрации / Программные системы: Теория и приложения. 2011.
- Амелькин С., Понизовкин Д. Оптимальное распределение проектов при проведении экспертизы / Электронные библиотеки: перспективные методы и технологии, электронные коллекции. 2010.
- Амелькин С., Понизовкин Д. Оптимальное проведение экспертизы образовательных процессов / Телематика. 2010. С. 158.
- Понизовкин Д. Определение меры близости профилей субъектов в задаче коллаборативной фильтрации / Управление и оптимизация неголономных систем. 2011.

Планируемые публикации

- акт о внедрении системы прогнозиварония Института почтовой связи
- акт о внедрении методов контентных PC в сервис компании $it - aces.com$
- зарегистрированное ПО для контентных PC