Математическая модель задачи top-N для контентных рекомендательных систем

к.т.н. Амелькин С.А., Понизовкин Д.М.
ИПС РАН им. А.К.Айламазяна
sergey.a.amelkin@gmail.com, denis.ponizovkin@gmail.com

Аннотация. В данной статье рассматриваются контентные рекомендательные системы, решающие задачу top-N. Предлагается математическая модель контентной рекомендательной системы, основанная на нечетких множествах, критерий оценки качества рекомендаций и алгоритм решения задачи.

Ключевые слова: контент пользователя, контент объекта, расстояние между контентами, контентная рекомендательная система, задача top-N, критерий оценки качеств рекомендаций, нечеткие множества.

Рекомендательная система или сервис (далее РС) – это информационные системы поддержки принятия решений, целью которых является прогноз пользовательских оценок объектов (статьи, книги [1], кинофильмы [2], конкурсные работы и т.п.). Такой прогноз может быть сделан на основании информации о пользователях, объектах и предварительно введенных пользователями оценках. РС могут быть использованы как в коммерческих целях для поиска целевых групп пользователей и продвижения товара в этих группах, так и для организации экспертного оценивания, где РС позволяют существенно снизить нагрузку на экспертов.

В статье будут рассматриваться *контентные* РС [6], которые основаны на анализе информации о характеристиках объекта и пользователя (что и называется *контентом*), или о том, какие объекты были уже выбраны пользователем.

Рассмотрим задачу выделения N объектов с наивысшими прогнозными значениями для каждого пользователя. Такая задача получила название задачи рекоменжации top-N [14]. Во втором разделе производится постановка данной задачи, в третьем разделе описывается математическая модель, в четвертом вводится формальная постановка задачи с алгоритмом ее решения, пятый раздел посвящен практическим результатам.

Множество пользователей будем представлять с помощью множества идентификаторов $U = \{1, 2, \dots, u, \dots, K\}$. Множество объектов также будем представлять с помощью множества идентификаторов $T = \{1, 2, \dots, t, \dots, L\}$.

Каждому пользователю u, задавшему запрос на N рекомендаций, назначается подмножество объектов T_N, где $T_N \in P(T)$, $|T_N| = N$, $P(T)$ — булеан множества T. Назовем множество T_N *профилем* пользователя u длины N и обозначим как $prof(u, N)$. Профили не могут формироваться наугад, т.к. каждый пользователь имеет свои предпочтения, а каждый объект — стилистические направленности, и, таким образом, каждый объект соответствует предпочтениям пользователя в какой-то мере. Для определения этой меры будем пользоваться расстоянием d_{ut} между пользователем u и объектом t. Предположим, что

$$\forall u \; \exists t : d_{ut} = 0. \qquad (1)$$

Множества пар вида (u, t), созданных в результате рекомендаций, таких как $t \in prof(u, N)$, назовем *распределением Q*.

Задачу top-*N* поставим следующим образом: необходимо создать алгоритм, находящий такое распределение *Q*, что среднее расстояние по *Q* стремится к минимуму:

$$\overline{d_{ut}}^{Q} \to \min. \qquad (2)$$

Поскольку каждый пользователь имеет свои предпочтения, а объект — стилистические

направленности, введем такие множества, элементы которых описывают данную информацию. Введем множество семантик пользователей C_U, $|C_U| = n \cdot c_i^u$, $i = 1..n$, принимающих значения от 0 до 1 и множество семантик объектов C_T, $|C_T| = m \cdot c_i^t$, $i = 1..m$, принимающих значения от 0 до 1. $C_U \cap C_T = \emptyset$. *Контент* пользователя (или объекта) — это информация о семантиках пользователя (или объекта), которая может быть представлена, к примеру, в векторном виде [5, 6].

Значение семантики представляет степень принадлежности семантики к пользователю или объект, поэтому будем рассматривать контенты как нечеткие множества [17]. Контент пользователя будем представлять в виде нечеткого множества $\left\{\left(c_1^u \middle| \mu\left(c_1^u\right)\right), \ldots, \left(c_n^u \middle| \mu\left(c_n^u\right)\right)\right\}$, $c_i^u \in C_U$ где множество C_U — универсальное множество, $\mu\left(c_i^u\right)$ — функция принадлежности элемента c_i^u контенту; контент объекта – $\left\{\left(c_1^t \middle| \nu\left(c_1^t\right)\right), \ldots, \left(c_m^t \middle| \nu\left(c_m^t\right)\right)\right\}$, $c_i^t \in C_T$, где множество C_T — универсальное множество, $\nu\left(c_i^t\right)$— функция принадлежности элемента c_i^t контенту. Также введем две операции: *ucont* и *tcont* однозначно сопоставляющие идентификаторам пользователя и объекта их контенты.

В дальнейшем нам понадобятся три операции над контентами:

1. Объединение $A \cup B$ — наименьший контент, содержащий A и B:

$$\forall c \in C_U : \mu_{A \cup B}(c) = \max\left(\mu_A(c), \mu_B(c)\right), \ \forall c \in C_T : \nu_{A \cup B}(c) = \max\left(\nu_A(c), \nu_B(c)\right).$$

2. Пересечение $A \cap B$ – наибольший контент, содержащийся одновременно в контентах A и B:

$$\forall c \in C_U : \mu_{A \cup B}(c) = \min\left(\mu_A(c), \mu_B(c)\right), \ \forall c \in C_T : \nu_{A \cup B}(c) = \min\left(\nu_A(c), \nu_B(c)\right).$$

3. Два контента A и B дополняют друг друга ($B = \overline{A}$ или $A = \overline{B}$), если:

$$\forall c \in C_U : \mu_{A \cup B}(c) = 1 - \mu_B(c), \ \forall c \in C_T : \nu_{A \cup B}(c) = 1 - \nu_B(c).$$

Необходимо ввести расстояние между элементами множеств U и T. Для этого воспользуемся информацией о пользователях и объектах, то есть введем расстояние следующим образом:

$$d_{ut}(u,t) = d\left(ucont(u), tcont(t)\right). \qquad (3)$$

Однако множества $ucont(u)$ и $tcont(t)$ имеют разную размерность и являются подмножествами различных универсальных множеств.

Поэтому для введения расстояния d_{ut} надо:

1. ввести расстояние между подмножествами множества T.
2. ввести отображение $uast : U \to T$ такое, что $\forall u$, $t \in uast(u) : d_{ut} = 0$.

Введем расстояние d_t на множестве T между элементами t_1, t_2 такое, что оно удовлетворяет следующим свойствам:

- $d_t(t_1, t_2) \geq 0$, если $t_1 = t_2$
- выполнение условия $d_t(t_1, t_2) \geq 0 \to t_1 = t_2$ необязательно
- $d_t(t_1, t_2) = d_t(t_2, t_1)$
- $d_t(t_1, t_3) \leq d_t(t_1, t_2) + d_t(t_2, t_3)$ (неравенство треугольника).

Зададим расстояние следующим образом:

$$d_{t(t_1,t_2)} = f\left(tcont(t_1), tcont(t_2)\right),$$

где f — любая псевдометрическая функция.

Введем расстояние d_T на множестве T между его элементами A и B:

$$d_T(A,B)=\inf\{d_t(t_1,t_2)\},\ t_1\in A, t_2\in B, \quad (4)$$

Отображение $uast$ будем строить на основании информации о контентах: $uast(u)=T'$, где $\forall t\in T': tcont(t)\in uast_c(ucont(u))$. Отображение $uast_c$ является нечетким отображением, так как не всегда можно четко определить отношения между предпочтениями пользователя и стилистическими направленностями объекта. Таким образом, $uast_c(A)$ контента A при нечетком отображении $\psi: C_U\times C_T\to[0,1]$ называется контент B с функцией принадлежности $\nu=\max(\min(\mu,\psi(A,B)))$.

Свойства функции $uast_c$:

• $uast_c(A\cup B)=conv(uast_c(A)\cup uast_c(B))$– отображение выпуклой оболочки. Следует из определения объединения контентов (минимальный контент, содержащий оба контента):

• $uast_c(A\cap B)=\overline{\overline{uast_c(A)}\cup\overline{uast_c(B)}}$.

Значения функции $uast_c$ определяются функцией ψ, так как значения μ известны (это данные о пользователе). Данная функция может быть задана таблично экспертом. Также функция ψ может быть выведена из исходных данных: установленных зависимостей между харатеристиками пользователей и объектов путем сбора известной информации о том, какие объекты пользователь уже предпочел.

Назовем *единичным* контентом такой контент, значения функции принадлежности которого равны нулю для всех его элементов, кроме одного:

$$\psi\{(c_1|0),(c_2|0),\ldots,(c_i|\mu(c_i)),(c_{i+1}|0)\ldots\},$$

Определим исходные данные как множество единичных контентов. Остальные контенты можно выразить с помощью объединения/пересечения элементов исходных данных.

В случае, когда мы располагаем только историей выбора объектов пользователями, не имея исходных данных, их можно получить путем комбинации данных истории и операции пересечения. Но для этого нужно располагать объемной историей, чтобы можно было получить необходимые единичные контенты.

Мы ввели способ построения расстояния между пользователем u и объектом t, реализуемый в 2 шага:

1. построение расстояния на множестве объектов;

2. построение расстояния между пользователем и объектом с использованием функции отображения $uast$, т.е. расчет расстояния между подмножеством $uast(u)$ и объектом t [4].

Приведем пример с известной функцией расстояния из информационного поиска – косинус угла между векторами $\cos(u,t)$ [17]:

1. $\cos'(t',t)=\sum_{i=1}^{m}(c_i^t)(c_i^{t'})\Big/\sqrt{\sum_{i=1}^{m}(c_i^t)^2\sum_{i=1}^{m}(c_i^{t'})^2}$;

2. $\cos(u,t)=\inf\cos'(t',t),\ t'\in uast(u)$.

Лемма:

Если существует мера сходства $f(t',t): T\times T\to R$ и отображение $uast$, то можно определить функцию расстояния $d_{ut}: U\times T\to R$ [3].

$$d_{ut}=\inf\{f(t',t): t'\in uast(u)\}, \quad (5)$$

причем, если f является псевдометрикой, то d_{ut} будет также псевдометрикой. Если

отображение *uast* не введено, то всегда можно ввести тождественное преобразование.

Критерием задачи top-N является значение среднего расстояния d_{ut} по распределению. В терминах, введенных в вышеописанных разделах, формально задачу можно записать так:

$$ed(Q)=\frac{1}{|Q|}\sum_{(u,t)\in Q}\frac{1}{N_u}\sum_{t\in prof(u,N_u)}d_{ut}(u,t), \qquad (6)$$

где:

- ed — обозначение критерия;
- Q — распределение;
- N_u — количество объектов, рекомендованных пользователю u.

Обозначим Q^* решение задачи $ed(Q^*)\to\min$. Задачей top-N назовем нахождение такого распределения Q^*.

Зачастую в качестве решения задачи top-N используются следующие критерии:

- точность[10, 13];
- полнота[9, 12];
- map [9, 12];
- DCG [9, 18];
- $nDCG=\frac{DCG}{IDCG}$ [9, 12, 13, 18];
- точность для N объектов $p@N$ [6, 10, 19].

Утверждение:

Если выполняются условия леммы и решение задачи top-N было получено на основе использования расстояния между пользователем и объектом, то критерий такого решения является частным случаем критерия ed.

Данное утверждение следует из основного свойства критериев: все они монотонно зависят от значения функции расстояния между пользователем и объектом.

Так как алгоритм решения данной задачи не может быть найден аналитически, то алгоритм решения был построен на основе метода имитации отжига [4].

Обозначения и определения, принятые в алгоритме:

- $rand(Q)$ – выбор случайного распределения;
- $rand(0,1)$ – выбор случайного числа в интервале $0,1$;
- $neighbour(Q)$ – выбор *соседнего* распределения.

Соседним распределением Q' к распределению Q назовем такое распределение, которое отличается от Q одним элементом одного профиля, сохраняя длину профиля:

- return Q — результатом работы алгоритма является распределение Q;
- *соседним распределением* Q' к распределению Q назовем такое распределение, которое отличается от Q одним элементом одного профиля, сохраняя длину профиля.

Алгоритм представляет собой стохастическую модификацию алгоритма спуска, в котором переход на каждом шаге является достоверным, если новое положение лучше предыдущего, и случайным в противном случае. Вероятность случайного перехода в худшее положение убывает в течение времени. Пошаговое описание алгоритма:

1. Инициализация

a. $i=0$;

b. $rand(Q)$: случайным образом выбираем исходное распределение;

c. $threshold=\varepsilon$: фиксируем точность алгоритма;

d. Присваиваем коэффициентам r и Δ начальные значения, в данном алгоритме Δ

характеризует интенсивность движения при ухудшении положения, r – скорость уменьшения Δ.

2. Вычисление

a. $Q' = neighbour(Q)$;

b. Вычисляем $ed(Q)$.

3. Переходы

a. **Если** $|ed(Q) - ed(Q')| < treshold$ **то** переходим к шагу 4;

b. **Иначе: Если** $ed(Q') < ed(Q)$ **то** $Q = Q'$: поскольку мы ищем минимальное значение критерия $ed(Q)$, то при уменьшении значения критерия переход к новому распределению является достоверным;

c. **Иначе: Если** $ed(Q') - ed(Q) < -\Delta \ln(rand(0,1))$, **то** $Q = Q'$: при увеличении значения критерия переход к новому распределению звисит от реализации случайной величины $rand(0,1)$.

d. $\Delta = \Delta \cdot r$; перейти к шагу 2;

4. Вывод результатов

a. Выводим значение распределения Q, алгоритм завершает работу.

Для оценки приведенного метода за основу были взяты исследования, опубликованные в статье [5]. В данной статье рассматривается задача рекомендации top-N, которая решается на базе семи различных моделей PC, основанных на семи различных мерах сходства(для данной статьи были взяты четыре меры сходства). Результаты решения сравниваются на основании оценочных мер, которые будут приведены ниже. Для сравнения данные модели тестировались на базе данных LastFm по следующей методологии: создавались обучающие выборки, содержащие 80% от профиля каждого пользователя (где профиль — это множество объектов, уже выбранных пользователем), остальные 20% считались закрытыми данными, используемыми для тестирования моделей. Ниже в таблице приведены опубликованные в [5] результаты тестов по базе данных LastFm:

Таблица 1

Результаты исследований статьи на данных LastFm [5]

Мера сходства	P@5	P@10	P@20	MAP	NDCG
Tf	0.028	0.021	0.014	0.011	0.085
Cos	0.234	0.109	0.059	0.041	0.202
Tf-idf	0.292	0.221	0.144	0.115	0.350
Bm25	0.226	0.105	0.075	0.051	0.216

Приведенные в [5] меры сходства рассматривались в терминах данной статьи в качестве $d_T :: T \times T \to R$ и использовалось тождественное преобразование $uast(u) = u$. Задачи решалась с помощью алгоритма, описанного выше.

Таблица 2

Результаты работы SAACT на данных LastFm

Мера сходства	P@5	P@10	P@20	MAP	NDCG
Tf	0.572	0.562	0.532	0.428	0.580
Cos	0.384	0.384	0.354	0.265	0.469
Tf-idf	0.314	0.276	0.208	0.170	0.370
Bm25	0.252	0.228	0.206	0.080	0.296

Литература

1. Linden G., Smith B., York J. Amazon.com Recommendations Item-to-Item Collaborative Filtering, Internet Computing, IEEE, vol. 7, pp 76 – 80.

2. R. Bell, Y. Koren, C. Volinsky (2007). The BellKor solution to the Netflix Prize.
3. Келли Дж. Общая топология. – М: Наука, 1968. 384 с.
4. Kirkpatrick S., Gelatt C.D., Vecchi M.P. Optimization by simulated annealing. Science. v. 220 (1983), pp 671–680.
5. Cantador, Iván and Bellogn, Alejandro and Vallet, David, Content-based recommendation in social tagging systems. ACM RecSys '10, pp. 237-240, 2010
6. Adomavicius G., Tuzhilin A. Toward the next generation of recommender systems: a survey of the state-of-the- art and possible extensions // IEEE Transactions on Knowledge and Data Engineering, vol. 17, no. 6, pp. 734-749, 2005.
7. Xiaoyuan Su, Taghi M., Khoshgoftaar A survey of collaborative filtering techniques // Advances in Artificial Intelligence Volume 2009, January 2009.
8. Burke R. Hybrid recommender systems: survey and experiments // User Modelling and User-Adapted Interaction, vol. 12, no. 4, pp. 331–370, 2002.
9. Baeza-Yates, R., Ribeiro-Neto, B. 1999. Modern Information Retrieval. Addison Wesley.
10. Andrew I. Schein, Alexandrin Popescul, Lyle H. Ungar, David M. Pennock. Methods and Metrics for Cold-Start Recommendations (2002). Proceedings of the 25th Annual International ACM SIGIR Conference
11. Miha Grchar and Dunja Mladenich and Marko Grobelnik Data sparsity issues in the collaborative filtering framework.Proceedings of the WebKDD'05 Proceedings of the 7th international conference on Knowledge Discovery on the Web: advances in Web Mining and Web Usage Analysis, pp58-76 Pages 58-76
12. Herlocker J.L., Konstan J.A., Terveen L.G., Riedl J.T. Evaluating collaborative filtering recommender systems//ACM Transactions on Information Systems, vol.22, no.1, pp. 5–53, 2004
13. Амелькин С.А. Оценка эффективности рекомендательных систем // Электронные библиотеки: перспективные методы и технологии, электронные коллекции. 2012.
14. Deshpande M., Karypis G. Item-based top-N recommendation algorithms // ACM Transactions on Information Systems, vol. 22, no. 1, pp. 143–177, 2004.
15. Кофман А. Введение в теорию нечетких множеств., М: Радио и связь, 1982
16. Заде Л.А. Тени нечетких множеств //Проблемы передачи информации, т. 2, вып. 1, март N. 2., с 37-44
17. Маннинг К.Д., Рагхаван П., Шютце Х. Введение в информационный поиск // Вильямс. М.: 2011. 528 с
18. Evaluating Recommender System // Guy Shani and Asela Gunawardana / Novermber 2009. URL: http://research.microsoft.com/pubs/115396/evaluationmetrics.tr.pdf

Сравнительная характеристика экологичности углеводородных растворителей и силиконового растворителя D_5 для процесса химической чистки.

Чл.- корр. РАН д.т.н. проф. Систер В.Г., ГрищенкоА.А., к.б.н. доц. Миташова Н.И., к.х.н. Баланова Т.Е.*

Университет машиностроения
anna400371@mail.ru, mitanieko@mail.ru
**ОАО «ЦНИИБЫТ»*
8(499)1860334

Аннотация. В статье изложены основные эколого-токсикологические характеристики нефтяного растворителя КВЛ и альтернативных – Солвон К-4 и силиконового D_5, используемых в настоящие время в химической чистке одежды. Отмечается наибольшая токсичность по фитотесту растворителя Солвон К-4.