ISSN 2079-3316 ПРОГРАММНЫЕ СИСТЕМЫ: ТЕОРИЯ И ПРИЛОЖЕНИЯ № 4(8), 2011, с. 107–114

УДК 51-7

Д. М. Понизовкин

Построение оптимального графа связей в системах коллаборативной фильтрации

Аннотация. Методы коллаборативной фильтрации — это множество алгоритмов и методик, позволяющих производить прогноз оценок пользователя информационных систем, основываясь на оценках других пользователей, что позволяет помогать людям в выборе и анализе огромного количества интересующей их информации. В статье рассмотрен способ построения функции расстояния между пользователями системы и объектами (которым пользователи ставят оценки), основываясь на функции расстояния между пользователями. Также в статье приведен критерий качества прогнозов.

Ключевые слова и фразы: коллаборативная фильтрация, рекомендательные системы, критерий прогнозирования, расстояние между пользователем и объектом.

Введение

Повседневно люди получают огромное количество информации от собеседников, газет, новостей, Интернета и т.д. Для того, чтобы упростить процесс анализа большого количества информации, например, при выборе книги, были созданы *рекомендательные системы (далее РС)*.

Разработчики одной из первых РС [1] ввели термин, который широко применяется на данный момент, — это термин *коллаборативная фильтрация* или *collaborative filtering* (далее КФ). В статье [1], где был введен данный термин, описывается экспериментальная система электронных писем, создание которой мотивировалось ростом электронных писем, наводняющих пользователя, большинство из которых пользователю не интересны. Целью данной системы являлась фильтрация ненужных писем. Помимо фильтрации по содержанию

Работа проводилась при финансовой поддержке Министерства образования и науки Российской Федерации.

(*content-based filtering*), система поддерживала КФ. Данная технология обозначает, что различные люди коммуницируют друг с другом, помогая отфильтровывать сообщения, записывая реакцию на прочитанные документы.

Эта техника часто используется при создании РС, чье предназначение состоит в рекомендации различных объектов пользователям таких систем. Популярной РС является веб-сервис Amazon. В нем используется база данных предпочтений пользователей по отношению к различным объектам для создания рекомендаций пользователям.

В общем виде сценарий КФ можно описать так [2]:

существует множество пользователей $\{u_1, u_2, \ldots, u_n\}$ и множество объектов $\{o_1, o_2, \ldots, o_m\}$. Каждый пользователь имеет список оцененных им объектов. Оценки могут принадлежать разным шкалам от 1 до 10, от 1 до 5 и т.д., а также разным типам шкал [3]: порядковой или относительной. Если пользователь u_i желает получить рекомендацию (или прогноз своей оценки на неоцененный объект), то по известным оценкам устанавливаются ближайшие по предпочтениям (или по оценкам на одни и те же объекты) пользователи к u_i. Далее система выдает рекомендации пользователю u_i (или рассчитывает прогнозную оценку на объект), исходя из оценок ближайших к u_i пользователей по предпочтениям.

Стандартные шаги данных алгоритмов КФ можно описать так:

- вычисление схожести между двумя пользователями,
- вычисление прогноза путем взятия среднего взвешенного ото всех оценок пользователя на прогнозируемый объект.

Если задача заключается в том, чтобы сгенерировать некоторое число топовых рекомендаций для пользователя u_i, тогда в качестве топовых избираются объекты, которые чаще всего оценивались ближайшими по предпочтениям пользователями к u_i.

В данной статье предложена модель прогнозирования, основанная на расстоянии между пользователями и объектами: каждый пользователь системы запрашивает определенное число объектов. Исходя из информации об объектах и пользователях, система производит выбор объектов пользователю, которые понравятся пользователю (или на которые он поставит высокую оценку).

1. Данные

Дано два множества: множество пользователей U и объектов O. Информация о пользователях и объектах представлена с помощью характеристик. Пользователям соответствует множество характеристик C_u, объектам — C_o. На множествах C_u и C_o введены порядковые отношения (то есть элементы этих множеств являются ранговыми величинами, которые можно только сравнивать). В общем случае размерность множеств не совпадает, и значения элементов разных множеств принадлежат разным порядковым шкалам.

Информацию о пользователях и объектах можно представить векторами характеристик, то есть описать элементы множеств U и O как

$$u \in U = (c_1^u, \ldots, c_N^u), \quad \text{где } c_i^u \in C_u, \quad i = \overline{1..N}, \quad N = |C_u|;$$
$$o \in O = (c_1^o, \ldots, c_M^o), \quad \text{где } c_j^o \in C_o, \quad j = \overline{1..M}, \quad M = |C_o|.$$

В общем случае каждая характеристика пользователя отображает отдельную область знаний, объекта — стилистическую направленность. Между характеристиками пользователей и объектов можно построить соответствие. Например, в системе экспертного оценивания каждого пользователя (эксперта) можно описать вектором, в котором каждая характеристика соответствует некоторой научной области, а ее значение — степени компетенции эксперта в этой области. Объект (проект) можно описать вектором, в котором каждая характеристика соответствует некоторой научной тематике, а ее значение — степени принадлежности темы проекта к тематике. Между такими характеристиками нетрудно построить соответствие. Другой пример: в системе http://www.last.fm/, вырабатывающей музыкальные рекомендации, областью знаний пользователя является его музыкальный вкус, стилистической направленностью объекта — музыкальный стиль.

2. Расстояние от пользователя до объекта

Пусть известно отображение $f : O \to U$, которое можно построить через установление соответствия между характеристиками пользователя и объекта. Данная функция отображает стилистическую направленность объекта o на область интересов пользователя u. Будем считать, что одному элементу множества O соответствует некоторое подмножество множества U. Для каждой ИС отображение $f : O \to U$

может быть реализовано по-разному, так как зависит от наборов характеристик и шкал, выбранных для них.

Как говорилось выше, каждый элемент множества U описывается вектором его характеристик. Зададим значимость характеристик с помощью шкалы весов. В общем случае будем задавать веса по следующему правилу: *шкале характеристик* $S = (s_1, \ldots, s_k)$, *где* s_k — *максимальное значение характеристики и* $s_i \geq s_j, i > j$, *соответствует шкала весов* $W = (w_1, \ldots, w_k)$, *где* $w_k = 0$, $\forall i > j : w_i \leq w_j$, $\forall i : w_i \geq 0$ *и* $w_i \in \mathbb{R}$. Вес w_i соответствует значимости характеристики s_i. Например, шкале $S' = (0, 1, 2, 3)$ может соответствовать шкала весов $W' = (1, 0.7, 0.5, 0)$. Объекту o, описанному вектором характеристик $(2, 3, 0, 1)$, будет соответствовать вектор весов $(0.5, 0, 1, 0.7)$.

Установив веса для характеристик, введем расстояние d_U для элементов a, b множества O. Пусть $w^a = (w_1^a, \ldots, w_M^a)$ — вектор весов для объекта a, $w^b = (w_1^b, \ldots, w_M^b)$ — вектор весов для объекта b.

$$d_U(a,b) = \begin{cases} 0, \text{ если } \exists i = \overline{1..M} : w_i^a = w_i^b = 0; \\ 0, \text{ если } \forall i = \overline{1..M} : w_i^a = w_i^b; \\ \sum\limits_{i=1}^{M} (w_i^a + w_i^b) \text{ в остальных случаях.} \end{cases} \tag{1}$$

Утверждение 2.1. *Функция* d_U *обладает псевдометрическими свойствами [4].*

Доказательство.

(1) $a = b$. Из определения расстояния следует, что $d_U(a,b) = 0$, так как их вектора весов для a и b будут совпадать. Однако из условия $d_U(a,b) = 0$ не следует равенство элементов a и b (см. первое условие в определении функции расстояния).

(2) $d_U(a,b) = d_U(b,a)$. Если $d_U(a,b) = 0$, то и $d_U(b,a) = 0$, что следует из определения. Если $d_U(a,b) \neq 0$, то $d_U(a,b) = \sum\limits_{i=1}^{M} (w_i^a + w_i^b) = \sum\limits_{i=1}^{M} (w_i^b + w_i^a) = d_U(b,a)$.

(3) $d_U(a,b) \leq d_U(a,c) + d_U(b,c)$. Обозначим через w^a, w^b и w^c вектора весов для объектов a, b и c соответственно. Неравенство треугольника $\sum\limits_{i=1}^{M} (w_i^a + w_i^b) \leq \sum\limits_{i=1}^{M} (w_i^a + w_i^c) + \sum\limits_{i=1}^{M} (w_i^b + w_i^c)$ следует

из того, что $2 \cdot \sum_{i=1}^{M} w_i^c \geq 0$, так как веса являются неотрицательными величинами.

Множество пользователей является псевдометрическим множеством. □

Введем функцию расстояния между двумя множествами U и O. Расстояние от точки x до подмножества A относительно псевдометрики d определим как $D(A, x) = \inf\{d(x, y) : y \in A\}$ [4].

Таким образом, можем ввести функцию расстояния от пользователя до объекта:

$$d_{UO}(u \in U, o \in O) = \inf\{d_U(x, y) : y \in f(o)\}. \tag{2}$$

3. Модель прогнозирования и ее представление с помощью графа

Результатом работы рекомендательной системы или системы КФ является некоторое множество, состоящее из пар вида (пользователь × объект) (в случае с прогнозными оценками паре будет соответствовать число — спрогнозированная оценка пользователя).

Результат работы таких систем можно представить с помощью графа связей. *Прогнозным графом связей* (далее граф связей) назовем такой граф $g = (V = \{U, O\}, E)$, вершины которого являются элементами из множеств O и U, ребра представляются парой $(u \in U, o \in O)$. Ребро, соединяющее элемент $u \in U$ с элементом $o \in O$ означает то, что объект $o \in O$ был рекомендован (или на него была спрогнозирована оценка) объекту $u \in U$. Каждое ребро имеет вес, который соответствует значению функции расстояния $d_{UO}(u \in U, o \in O)$. Из каждого элемента $u_i \in U$, $i = \overline{1..N}$, можно провести L_i $(i = \overline{1..N})$ ребер к L_i разным объектам. Число L_i соответствует количеству запрошенных пользователем рекомендаций (или прогнозов).

Тем самым задачу выработки прогнозирования можно описать как нахождение такого графа связей, у которого сумма весов ребер стремится к минимуму. Иными словами, рекомендовать такие объекты для пользователям, чтобы среднее расстояние между ними и пользователями было как можно меньше.

Введем функцию $e : G \to \mathbb{R}$

$$e(g \in G) = \sum_{i=1}^{N} \sum_{j=1}^{L_i} d_{UO}(u_i, o_{ij}). \tag{3}$$

Данная функция ставит в соответствие графу связей число, равное среднему расстоянию между объектами и пользователями. Значение этой функции является показателем эффективности прогноза. Чем оно меньше, тем прогноз лучше.

Задача: требуется найти минимум функции $e(g \in G)$ на множестве G.

Решение задачи: поставленная задача может быть решена с помощью полного перебора всего множества графов G. Но это решение требует большого времени вычисления. В данной работе предложен алгоритм, основанный на методе имитации отжига (*simulated annealing*) [5].

Определение 3.1. Профилем *пользователя* u_i, $i = \overline{1..L_i}$, *назовем* $O' \subset O = o_1, \ldots, o_{L_i}$, *такое, что* $\forall j = \overline{1..L_i}\, \exists e_j \in E = (u_i, o_j)$ *(иными словами, профиль пользователя — множество назначенных ему объектов).*

Определение 3.2. Соседним графом *к графу* g *считается такой граф* h, *в котором существует только один профиль пользователя, отличающийся от профиля того же пользователя в графе* g, *на один объект.*

Далее с помощью псевдокода описывается алгоритм нахождения минимума функции $e(g \in G)$.

- **Шаг 0**. Инициализация алгоритма.
  - Инициализируем переменную $i := 0$.
  - Инициализируем переменную $t := T$ — начальная температура. Для разных систем может быть разной.
  - Выберем случайный граф $g \in G$.
  - $r \in (0, 1)$ — константа, нужная для понижения температуры.
- **Шаг 1**. Если $t = 0$, алгоритм заканчивается. Решением является граф g.
- **Шаг 2**. Рассчитаем значение $e(g) = val$.
- **Шаг 3**. Выберем соседний к g граф g'.
- **Шаг 4**. Рассчитаем значение $e(g') = val'$.
- **Шаг 5**. Если $(val > val')$:

(1) $g := g'$ (текущим графом считаем граф g'),
(2) $t := T$.

Иначе:

(1) рассчитаем вероятность P перехода в граф g' и берем случайное число $p \in (0, 1)$;
(2) если $p > P$, то $g := g'$(текущим графом считаем граф g');
(3) иначе текущим графом является g.

4. Практические результаты

Для ИС http://edu.botik.ru/ был написан модуль, реализующий описанный в работе алгоритм. Алгоритм был протестирован на результатах, полученных на XV молодежной научной конференции «Наукоемкие информационные технологии» г. Переславля.

Пользователями системы являются студенты и лектора, объектами — научные проекты, представленные участниками конференции. После распределения научных проектов пользователи системы выставляют оценки полученным проектам. В результате оценивания определяются победители конференции. Каждый пользователь системы описывается вектором характеристик, соответствующих научным областям. Значение характеристики пользователя определяет степень компетентности пользователя в научной области. Проекты также описываются векторами характеристик, которые соответствуют научным тематикам. Значение характеристики объекта определяет степень его принадлежности научной тематике. На основании значений векторов производится распределение. Было получено значение функции показателя эффективности прогноза, равное 17169. Значение данной функции на графе связей, полученном при работе алгоритма, использующегося ранее в системе — 40766. Таким образом, получили улучшение более, чем в два раза.

5. Заключение

Предложена модель прогнозирования в системах КФ и способ его оценивания. Данная модель не нуждается в последовательном переборе пользователей и нахождении для каждого из них множества ближайших пользователей по вкусам, что ведет к сокращению расчетов. Предложен расчет близости между пользователями, основанный на псевдометрических оценках, а также иной критерий, который

применим практически. Описанные методы реализованы и показана их практическая эффективность.

Список литературы

[1] Goldberg D., Nichols D., Oki B., Terry D. *Using collaborative filtering to weave an information* // Communications of ACM, 1992. **35**, no. 12, p. 61–70 ↑[]

[2] Heckerman D., Breese J.S., Kadie C. *Empirical Analysis of Predictive Algorithms for Collaborative Filtering* : // Proceedings of the 14th Conference on Uncertainty in Artificial Intelligence, 1998 ↑[]

[3] Орлов А. И. Прикладная статистика. М. : Изд-во «Экзамен», 2004. — 656 с. ↑[]

[4] Келли Д. Общая топология. М. : Изд-во «Наука», 1968. — 384 с. ↑2.1, 2

[5] Kirkpatrick S., Gelatt C., Vecchi M. *Optimization by Simulated Annealing* // Communications of ACM, 1988. **220**, no. 4598, p. 671–681 ↑3

D. M. Ponizovkin. *Construction of optimal graph of relations between sets of subjects and objects in collaborative filter systems.*

Abstract. Collaborative filtering (CF) is the process of filtering for information or patterns using techniques involving collaboration among multiple agents, viewpoints, data sources, etc. Applications of collaborative filtering typically involve very large data sets. In article presents metric function between set of users and set of items and evaluating method of predictions.

Key Words and Phrases: collaborative filtering, recommender systems, evaluating of predictions, distance between user and item.

Образец ссылки на статью:

Д. М. Понизовкин. *Построение оптимального графа связей в системах коллаборативной фильтрации* // Программные системы: теория и приложения : электрон. научн. журн. 2011. № 4(8), с. 107–114. URL: http://psta.psiras.ru/read/psta2011_4_107-114.pdf