Во второй задаче соревнования была предостав-

лена статистика поведения пользователей, одна-

ко, не полная, как это принято в задачах такого

типа (кто, что и когда смотрел). Причиной такой

неполноты является одна из тенденций в раз-

витии современной IT-индустрии: обеспечение

анонимности действий пользователей в Интер-

нете. В последнее время актуальнейшей задачей

становится проблема правильного «огрубления»

информации, при котором задачи анализа данных

можно решать с приемлемым качеством, но нель-

зя восстановить исходные данные (подробные

логи пользователей). Организаторы соревнования

для такого огрубления применили метод пост-

троечных последовательностей (это наш перевод

термина «pooled sequences»).

Здесь корень извлекается поэлементно. В каче-

стве рекомендации алгоритм дает лекции, которые

соответствуют наибольшим значениям коорди-

нат в векторе . Вид решения опреде-

ляется технологией решения прикладных задач

«LENKOR», развиваемой автором, основные этапы

применения которой:

1. Выделение различных видов информации,

описание способов вычислений близости по каж-

дому виду.

2. Формирование линейной комбинации функ-

ций близости, настройка коэффициентов (методом

покоординатного спуска). При этом автоматически

определяются «ненужные» виды информации (соот-

ветствующие слагаемые входят с нулевыми весами).

3. «Деформирование» комбинации (попытка по-

строить нелинейную формулу решения путём перебо-

ра различных алгебраических выражений), настройка

коэффициентов (методом покоординатного спуска).

На третьем этапе приходится иметь дело с относи-

тельно небольшими выражениями (часть функций

близости отсеивается на втором этапе), что обеспе-

чивает возможность перебора нелинейных выраже-

ний различных типов. В данной задаче оптималь-

ным выражением оказалась линейная комбинация

корней, которая, впрочем, не сильно улучшает ли-

нейное решение (на 2%). Отметим, что значения ко-

эффициентов в формуле (5) и вид выражения ха-

рактерен для решения конкретной задачи (даже при

накоплении статистики в рассматриваемой задаче

оптимальные значения коэффициентов могут изме-

ниться). Здесь приводится точная формула для воз-

можности верификации полученных результатов.

В решение, которое автор выложил на сайт [3],

внесено ещё одно изменение: вектор

преобразуется в вектор

,

где *tj* – время выкладывания на сайт *j*-й новой лек-

ции, *tmin* – минимальное среди всех этих времён

новых лекций, *tmax* – максимальное (вычислялось в

днях). Такая поправка учитывает условия соревно-

вания: поскольку оценка происходит по имеющей-

ся статистике (т.е. по числу пользователей, посмо-

тревших новую лекцию после просмотра лекции из

контрольной выборки), то важна не только попу-

лярность лекции, но и сколько она была доступна

для просмотра. В *табл. 1* представлена зависимость

от .

Описанные алгоритмы достаточно просты, уни-

версальны, допускают возможности распаралле-

ливания. Решение получается в удобном виде: как

вектор оценок. Для рекомендации некоторого ко-

личества лекций достаточно отобрать столько наи-

больших элементов вектора, но, в принципе, парал-

лельно получаются оценки популярности каждой

лекции. Кроме того, алгоритмы такого типа могут

быть использованы в рамках алгебраического под-

хода [13] для формирования выражений над алго-

ритмами. По сути, в основе технологии «LENKOR»

лежат идеи алгебраического подхода: выбирается

«правильная» база пространства векторов оценок,

а затем настраивается алгебраическое выражение.

Предложенные методы могут использоваться для

других постановок задач. Например, алгоритм ре-

шения задачи «холодный старт» может быть легко

приспособлен к решению задач кредитного скорин-

га и оценки перспективности проектов. Отметим,

что для этих задач технология «LENKOR» и была

изначально разработана, хотя в задачах кредитно-

го скоринга она может уступать в эффективности

«стандартным» методам, например случайному лесу

(Random Forest). Алгоритм решения задачи «пост-

троечные последовательности» может быть исполь-

зован не только при построении рекомендательных

систем, но и в задачах прогнозирования *k*-значных

временных рядов (для поиска закономерностей типа

«три события определяют будущее событие») и ав-

томатической классификации текстов (в настоящее

время автор модифицирует этот метод для класси-

фикации на основе троек терминов).

Прочел понял что используется явный подход при сборе информации лекциях если не ошибаюсь…

Внедрено как система «LENKOR»…

Недостатков не выявил… Потому что еще не все понимаю наверно…