УДК 004.03

А.В. Щербаков, В.С. Нарышкин

Харьковский национальный экономический университет, Харьков

ИНТЕЛЛЕКТУАЛЬНАЯ ОБРАБОТКА ИНФОРМАЦИИ В СРЕДЕ WEB 2.0 НА ПРИМЕРЕ РЕШЕНИЯ ЗАДАЧИ ВЫРАБОТКИ РЕКОМЕНДАЦИЙ

В статье рассмотрен процесс интеллектуальной обработки информации, генерируемой пользователями Интернет, для решения задачи выработки рекомендаций. Приведены алгоритмы определения коэффициента подобия людей с помощью коэффициента Танимото, евклидового расстояния и коэффициента корреляции Пирсона. Предложен метод определения подобия с помощью классификационных тегов.

Ключевые слова: *информация, интеллектуальная обработка информации, коллаборативная фильтрация, Web 2.0, Интернет, выработка рекомендаций, подобие, коэффициент подобия, тег.*

Введение

Средства эффективной обработки информации в Интернете еще никогда не были настолько важны и востребованы, как сегодня.

Количество людей, осознанно или случайно пополняющих Интернет данными, постоянно растет. Они уже создали колоссальный массив данных, анализируя который можно многое узнать о пользователях – как они работают, что предпочитают, какие товары их интересуют, – да и вообще о человеческом поведении.

Разработанные методы коллаборативной фильтрации [1] позволяют дать автоматические прогнозы относительно интересов пользователя по собранной информации о вкусах других пользователей. В век информационного взрыва такие технологии могут оказаться чрезвычайно полезными, поскольку количество объектов даже в одной категории (такой, как фильмы, музыка, книги, новости, веб-сайты) стало настолько большим, что отдельный человек не способен просмотреть их все, чтобы выбрать подходящие.

Применение методов коллаборативной фильтрации в среде Web 2.0 [2] является следующим шагом на пути развития информационных технологий. В источнике [3] рассмотрены интеллектуальные приложения для Web 2.0, которые способны собирать и анализировать огромные массивы данных, имеющиеся в Сети.

Приведены алгоритмы машинного обучения и их применение в различных областях, а также способы, позволяющие с их помощью делать выводы из данных, полученных от многих людей.

Рассмотрены методы извлечения смысла из данных, в том числе и коллаборативная фильтрация, которая используется для рекомендования предметов Интернет сервисами.

Цель статьи – показать применимость методов коллаборативной фильтрации для выработки рекомендаций в среде Web 2.0.

Основная часть

Сущность Web 2.0 – это контент, генерируемый пользователями [2], но стоит учитывать тот факт, что тот контент, который пользователи вводят в Web 2.0 явно, составляет лишь малую часть всех генерируемых данных.

Способность собирать информацию и наличие вычислительных мощностей для ее интерпретации открыли новые возможности для интеллектуальных приложений.

Интеллектуальные приложения для Web 2.0 должны знать, как из этих данных добывать информацию – явно или путем анализа поведения пользователей в Интернете.

Собранная информация может быть использована для решения практических задач [3]. Выработка рекомендаций – одна из возможных областей применения полученной информации. Предпочтения некоторой группы людей можно использовать для того, чтобы рекомендовать что-то другим людям. У такой техники немало применений, в частности рекомендование товаров на сайте электронной торговли или помощь в отыскании нужной музыки и фильмов.

Информацию о предпочтениях можно собирать по-разному. Иногда данными являются купленные посетителем товары, а мнения об этих товарах представляются в виде голосования «да/нет» или оценки по пятибалльной шкале.

Собрав данные о том, что людям нравится, нужно как-то определить, насколько их вкусы схожи. Для этого вычисляется коэффициент подобия (или оценка подобия) людей (рис. 1). Существует несколько способов расчета коэффициента подобия:

коэффициент Танимото, евклидово расстояние и коэффициент корреляции Пирсона.

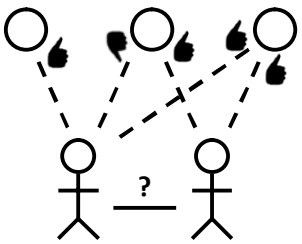

Рис. 1. Определение коэффициента подобия людей по их оценкам предметов

Коэффициент Танимото (1) описывает степень схожести двух множеств [4], применяется, в основном, для оценок «да/нет»:

$$k = \frac{c}{a + b - c}, \qquad (1)$$

где k – коэффициент Танимото (число от 0 до 1), чем он ближе к 1, тем более схожи множества;

a – количество элементов в первом множестве;

b – количество элементов во втором множестве;

c – количество общих элементов в двух множествах.

Оценка по евклидову расстоянию [5] – один из самых простых способов вычисления оценки подобия. В этом случае предметы, которые люди оценивали сообща, представляются в виде координатных осей. Теперь в этой системе координат можно расположить точки, соответствующие людям, и посмотреть, насколько они оказались близки:

$$p(x, y) = \sum_{k=1}^{n} \sqrt{(x_k - y_k)^2}. \qquad (2)$$

Расстояние, вычисленное по формуле (2), будет тем меньше, чем больше сходства между людьми. Однако нам нужна функция, значение которой тем больше, чем люди более похожи друг на друга:

$$p'(x, y) = \frac{1}{p(x, y) + 1}. \qquad (3)$$

Преобразованная функция (3) всегда возвращает значение от 0 до 1, причем 1 получается, когда предпочтения двух людей в точности совпадают.

Коэффициент корреляции Пирсона [6] – более сложный способ определить степень схожести интересов людей. Коэффициент корреляции – это мера того, насколько хорошо два набора данных ложатся на прямую.

Формула сложнее, чем для вычисления евклидова расстояния, но она дает лучшие результаты, когда данные плохо нормализованы, например, если некоторый критик устойчиво выставляет фильмам более низкие оценки, чем в среднем:

$$r_{xy} = \frac{n \cdot \sum_i x_i \cdot y_i - \sum_i x_i \cdot \sum_i y_i}{\sqrt{\left(n \cdot \sum_i x_i^2 - \left(\sum_i x_i\right)^2\right) \cdot \left(n \cdot \sum_i y_i^2 - \left(\sum_i y_i\right)^2\right)}}. \qquad (4)$$

Функция (4) возвращает значение от –1 до 1. Значение 1 означает, что два человека выставили каждому предмету в точности одинаковые оценки.

Предложенные формулы (1), (3), (4) могут использоваться для отыскания людей, которые имеют наилучшее соответствие с данным человеком. В целях повышения эффективности системы, можно заранее построить матрицу коэффициентов подобия всех пользователей P, определив подобие всех людей между собой. Это позволит использовать готовые значения на этапе рекомендования предметов, при этом экономя вычислительные ресурсы.

Найденные значения позволят рекомендовать предметы с учетом вкусов пользователей (рис. 2).

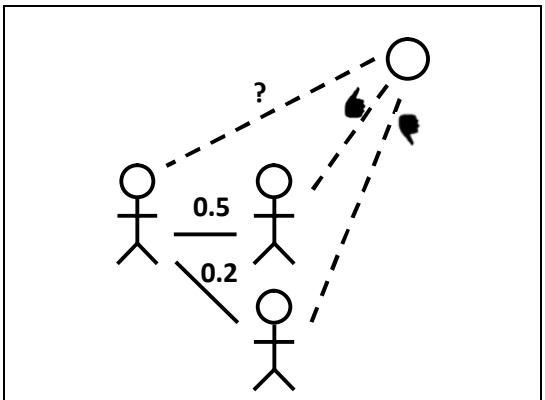

Рис. 2. Прогнозируемая оценка предмета, который пользователь не оценивал

Пользователи, которые имеют более схожие вкусы со вкусами данного пользователя, имеют большее влияние на прогнозируемую оценку предмета (5):

$$O_{x,y} = \frac{\sum_{i=1}^{n} P_{x,i} \cdot O_{i,y}}{\sum_{i=1}^{n} P_{x,i}}, \qquad (5)$$

где $O_{x,y}$ – оценка пользователем *x* предмета y;

P – матрица коэффициентов подобия пользователей.

Определив для пользователя его прогнозируемые оценки всех предметов, можно определить список предметов, которые наиболее вероятно ему понравятся.

Можно пойти дальше, присвоив каждому предмету классифицирующие слова (теги [7]), а также позволив пользователям самостоятельно при-

своить себе теги, основываясь на их увлечениях и предпочтениях (рис. 3). Это позволит учитывать тот факт, что специалист в определенной предметной области должен иметь больший вес своего голоса, чем остальные люди, а также то, что связь между двумя людьми, одинаково классифицировавших себя, определенно существует, даже если не была выявлена ранее.

Для определения оценки подобия объектов по классификационным тегам можно использовать предложенные выше формулы (1), (3), (4), заменив оценки на теги. Таким образом, чем больше встречается одинаковых тегов у объектов, тем более они подобны.

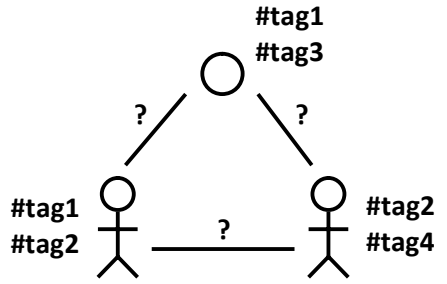

Рис. 3. Определение подобия объектов на основе классификационных тегов

Для оценки подобия пользователей можно использовать среднее взвешенное значение предложенных методов (6):

$$P' = \alpha \cdot P_o + (1-\alpha) \cdot P_\tau, \qquad (6)$$

где P_o – подобие пользователей, вычисленное по их оценкам предметов (рис. 1);

P_τ – подобие пользователей, вычисленное по классификационным тегам (рис. 3);

α – вес каждой составляющей, может принимать значение от 0 до 1.

Приняв во внимание схожесть между людьми и предметами, которые они оценивают (рис. 4), формула (5) принимает вид (7):

$$O_{x,y} = \frac{\sum_{i=1}^{n} P'_{x,i} \cdot O_{i,y} \cdot G_{i,y}}{\sum_{i=1}^{n} P'_{x,i} \cdot G_{i,y}}, \qquad (7)$$

где $O_{x,y}$ – оценка пользователем x предмета y;

P' – матрица коэффициентов подобия пользователей, найденная по формуле (6);

G – матрица коэффициентов подобия пользователей с предметами.

Стоит учесть то, что у определенных людей не будет совпадений по тегам с некоторыми предметами. Таким образом, их голос не будет учитываться при составлении рекомендаций.

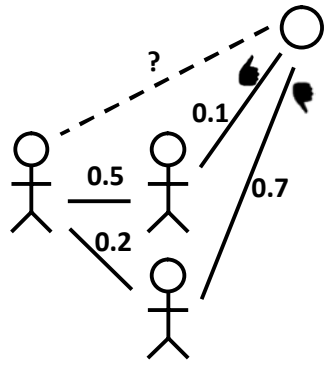

Рис. 4. Прогнозируемая оценка предмета, который пользователь не оценивал, с учетом подобия по классификационным тегам

Необходима формула (8), которая будет корректировать коэффициент подобия, чтобы учитывались даже те голоса, которые не имеют совпадений по тегам:

$$d' = \frac{d+k}{1+k}, \qquad (8)$$

где d' – скорректированный коэффициент подобия;

d – коэффициент подобия;

k – параметр модели.

Влияние параметра k на коэффициент подобия в формуле (8) показано на рис. 5.

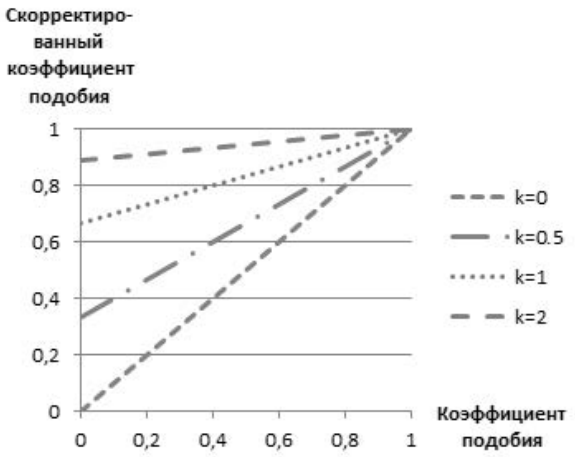

Рис. 5. Влияние параметра k на коэффициент подобия

Так как классификационные теги допускают возможность произвольного написания пользователями, это создает проблемы для сопоставления одинаковых по значению, но разных по написанию тегов.

Можно применить описанные до этого методы и для поиска подобных тегов: для этого нужно предметы заменить та теги, оценки пользователей заменить на 1, если тег принадлежит пользователю,

0 – если нет (рис. 6). Рассчитав подобие всех тегов по формулам (1), (3), (4), можно определить какие теги имеют одинаковый смысл.

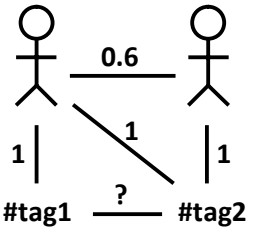

Рис. 6. Обнаружение взаимосвязей между тегами

Тот же самый подход может быть применен и для поиска предметов, похожих на заданный (рис. 7). Эта техника широко распространена в Интернет-магазинах для составления списка товаров, похожих на просматриваемый.

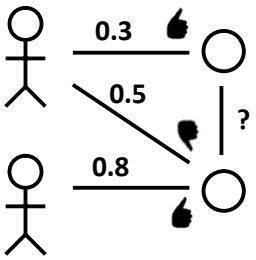

Рис. 7. Определение подобных предметов

Выводы

Рассмотрены принципы коллаборативной фильтрации в среде Web 2.0 для определения интересов пользователя на основе информации о вкусах других пользователей.

Приведены методы определения коэффициента подобия людей. Предложен метод определения подобия с помощью классификационных тегов.

Список литературы

1. *Коллаборативная фильтрация [Электронный ресурс]. – Режим доступа к ресурсу: http://ru.wikipedia.org/wiki/Коллаборативная_фильтрация.*

2. *Веб 2.0 [Электронный ресурс]. – Режим доступа к ресурсу: http://ru.wikipedia.org/wiki/Веб_2.0.*

3. *Сегаран Т. Программируем коллективный разум: пер. с англ. / Т. Сегаран. – СПб: Символ-Плюс, 2008. – 368 с., ил.*

4. *Использование коэффициента Танимото для поиска людей с одинаковыми предпочтениями [Электронный ресурс]. – Режим доступа к ресурсу: http://habrahabr.ru/blogs/algorithm/104901/*

5. *Евклидово пространство [Электронный ресурс]. – Режим доступа к ресурсу: http://ru.wikipedia.org/wiki/Евклидово_пространство.*

6. *Коэффициент линейной корреляции Пирсона [Электронный ресурс]. – Режим доступа к ресурсу: http://psystat.at.ua/publ/1-1-0-26.*

7. *Tag_(metadata) [Электронный ресурс]. – Режим доступа к ресурсу: http://en.wikipedia.org/wiki/Tag_(metadata).*

Поступила в редколлегию 21.02.2011

Рецензент: д-р физ.-мат. наук, проф. С.В. Смеляков, Харьковский университет Воздушных Сил им. И. Кожедуба, Харьков.

ІНТЕЛЕКТУАЛЬНА ОБРОБКА ІНФОРМАЦІЇ В СЕРЕДОВИЩІ WEB 2.0 НА ПРИКЛАДІ РОЗВ'ЯЗКУ ЗАДАЧІ ВИРОБЛЕННЯ РЕКОМЕНДАЦІЙ

О.В. Щербаков, В.С. Наришкін

У статті розглянутий метод інтелектуальної обробки інформації, яка генерується користувачами Інтернет, для розв'язку задачі вироблення рекомендацій. Наведені алгоритми визначення коефіцієнта подоби людей за допомогою коефіцієнту Танімото, евклідової відстані та коефіцієнту кореляції Пірсона. Запропонований метод визначення подоби за допомогою класифікаційних тегів.

Ключові слова: *інформація, інтелектуальна обробка інформації, колаборативна фільтрація, Web 2.0, Інтернет, вироблення рекомендацій, подоба, коефіцієнт подоби, тег.*

INTELLECTUAL PROCESSING OF THE INFORMATION IN THE WEB 2.0 ENVIRONMENT ON THE EXAMPLE OF THE PROBLEM DECISION OF RECOMMENDATIONS MAKING

O.V. Shcherbakov, V.S. Naryshkin

In article the method of intellectual processing of the information generated by users of the Internet, for the decision of a problem of recommendations making is considered. Algorithms of definition of people similarity factor using the Tanimoto coefficient, Euclidian distance and Pearson correlation coefficient are proposed. The method of definition of similarity by means of classification tags is offered.

Keywords: *information, intellectual processing of the information, collaborative filtration, Web 2.0, Internet, recommendations making, similarity, similarity factor, tag.*