УДК 004.42: 519.25

МЕТОД КЛАСТЕРИЗАЦИИ ТЕМАТИЧЕСКИХ ПРОФИЛЕЙ ПОЛЬЗОВАТЕЛЕЙ И ЕГО ПРИМЕНЕНИЕ ДЛЯ АНАЛИЗА ИНТЕРНЕТ-ТРАФИКА

Паутов К.Г., Попов Ф.А.

Бийский технологический институт (филиал), ФГБОУ ВПО «Алтайский государственный технический университет им. И.И. Ползунова», Бийск, e-mail: pautov@bti.secna.ru

В работе описывается метод кластеризации тематических профилей пользователей, построенных на основе анализа информации о посещенных ими информационных ресурсах в сети Интернет. Показано, что использование данной информации позволяет достаточно достоверно определить тематические предпочтения пользователей. В статье детально описаны исходные данные, процессы конструирования и отбора признаков, а также эксперименты с использованием метода агломеративной иерархической кластеризации и метода k-means. Полученные результаты позволяют судить об интересах и предпочтениях отдельных пользователей и групп, а также делать выводы о том, насколько целевым является использование ресурсов Интернет в организации. Информация о кластерах пользовательских предпочтений позволяет проводить оценку эффективности работы отдельных сотрудников, выявлять информационные потребности групп пользователей и обосновывать целесообразность вложений в развитие ИТ-инфраструктуры.

Ключевые слова: машинное обучение, кластеризация, тематический профиль пользователя, определение информационных потребностей

AN APPROACH FOR CLUSTERING OF THE THEMATIC USER PROFILES AND ITS APPLICATION FOR INTERNET TRAFFIC ANALYSIS

Pautov K.G., Popov F.A.

Biysk Technological Institute (branch) of the federal government budget of educational institutions of higher education «Altai State Technical University of I.I. Polzunov», Biysk, e-mail: pautov@bti.secna.ru

In this paper we described a method for clustering thematic user profiles that are based on analysis of information about their visits of information resources on the Internet. It is shown that the use of this information allows to reliably determining the thematic preferences of users. Initial data about Internet activity of users were obtained from the log files of the proxy server. The article describes in detail the process of input data preparation, feature engineering and construction of the feature matrix, as well as experiments using the agglomerative hierarchical clustering algorithm and k-means method. Moreover, we propose a method of constructing thematic profiles of users based on the data on the amount of traffic and frequency of visits web sites within thematic categories. Obtained results allow us to judge about the interests and information preferences of individual users and groups, and to draw conclusions about the effectiveness of the use of Internet resources in the organization. The clusters of user preferences allow assessing staff performance and feasibility of investment in the development of IT infrastructure.

Keywords: machine learning, clustering, thematic user profiles, analysis of thematic preferences

В связи с интенсивным развитием телекоммуникационных технологий, увеличением скоростей передачи данных и пропускной способности каналов связи все большую актуальность приобретает задача изучения информационных потребностей пользователей.

По оценке фонда «Общественное мнение» [7] сотрудник офиса ежедневно тратит 1,5 часа рабочего времени на решение личных вопросов, общение в социальных сетях, чтение новостей, посещение развлекательных порталов и т.д. В целом такое положение дел приводит к снижению эффективности использования рабочего времени сотрудниками организации, потерям времени, связанным с нецелевым использованием ресурсов Интернет и увеличению нагрузки на каналы связи [2, 6]. Вопросы, связанные с оптимизацией потребления Интернет-трафика, и методика поэтапного сокращения затрат на Интернет-трафик и потерь рабочего времени, связанных с нецелевым использованием ресурсов сети Интернет, детально описаны в [2].

В данной статье рассмотрено применение методов интеллектуального анализа данных (data mining) для оценки эффективности использования сотрудниками организации своего рабочего времени и ресурсов сети Интернет путем построения тематических профилей пользователей, определения их информационных потребностей, а также формирования кластеров пользователей со схожими тематическими предпочтениями.

Исходными данными для анализа является информация, полученная из журнальных файлов (log-файлов) прокси-сервера [4]. Данные о каждом обращении

пользователя к ресурсам Интернет обрабатываются прокси-сервером и сохраняются в виде текстового log-файла. Каждому запросу в log-файле соответствует одна строка.

Имея достаточно данных о посещенных пользователем веб-страницах, можно вполне успешно определить тематические категории, отражающие интересы и предпочтения данного пользователя, а также сделать вполне обоснованные предположения об эффективности его работы [1, 5]. Задача осложняется тем, что количество пользователей организации может исчисляться сотнями, а число посещенных ими веб-страниц – тысячами. При этом размеры сжатого дневного журнала прокси-сервера могут достигать 100 МБ и более. Для обработки и анализа таких объемов информации необходимо использовать автоматизированные методы, базирующиеся на применении машинного обучения.

Первоначальным этапом решения задачи data mining является процесс конструирования признаков (feature engineering). Это самый трудоемкий и одновременно самый ответственный этап, от которого напрямую зависит конечный результат всей работы. В нашем случае в качестве объектов выступают пользователи организации, а в качестве признаков – тематические категории посещенных ими веб-ресурсов. Такое признаковое описание позволяет формировать тематические профили пользователей. Тематический профиль пользователя – это векторное представление его интересов и тематических предпочтений, составленное на основе анализа посещенных веб-страниц. Совокупность тематических профилей пользователей образует матрицу, где каждая строка соответствует пользователю, а каждый столбец – признаку, в качестве которого выступает тематическая категория (рис. 1).

Значения признаков вычисляются исходя из частот обращения пользователя к ресурсам, входящим в каждую тематическую категорию, и объемов входящего трафика. Для улучшения качества будущей модели производят нормализацию значений признаков, приводя их к диапазону [0...1].

После окончания этапа конструирования признаков (feature engineering), как правило, производят отбор подмножества наиболее информативных и достоверных признаков для построения модели. Это позволяет снизить объемы обрабатываемой информации, избежать переобучения и в целом улучшить качество модели. В нашем случае будем группировать ресурсы по тематическим категориям, т.к. разумно предположить, что пользователи, интересующиеся одной тематикой, могут получать информацию из различных источников. Признаки, встречающиеся не более чем у одного пользователя, будем считать неинформативными.

В результате получим матрицу «пользователь/категория», состоящую только из информативных признаков. Как правило, матрица обладает большой размерностью, и при этом является разреженной (sparse matrix). Поэтому следует использовать реализации алгоритмов машинного обучения, способные работать с такими матрицами.

Для того чтобы получить первоначальное представление о кластерной структуре исследуемых данных, воспользуемся агломеративным алгоритмом иерархической кластеризации и разобьем исходное множество объектов на несколько кластеров. Для этого вычислим попарные расстояния между объектами, построив матрицу расстояний. В качестве меры расстояния воспользуемся расстоянием Евклида. На рис. 2 представлены гистограммы распределения расстояний между объектами.

Преобладание больших расстояний между объектами свидетельствует о неоднородности исследуемых данных. То есть информационные потребности пользователей в целом довольно сильно различаются.

← User_ID	← Наука/	← Развлечения/ Соц.сети/	← Семья/ Покупки/	← Компьютеры/	← Бизнес/ Финансы/	← СМИ/	← Развлечения/ Видеохостинги/
10.2.2.112	0.0	0.0	0.127390	0.208814	0.227901	0.0	0.196162
10.2.1.202	0.0	0.335914	0.0	0.0	0.0	0.0	0.310783
10.2.1.111	0.0	0.0	0.089418	0.0	0.0	0.124309	0.348838
10.2.12.244	0.0	0.630352	0.0	0.0	0.902977	0.0	0.0
192.168.10.22	0.0	0.0	0.0	0.751444	0.0	0.425993	0.344069
10.2.2.149	0.999673	0.0	0.0	0.0	0.0	0.114259	0.110382

Рис. 1. Тематические профили пользователей. Векторное представление

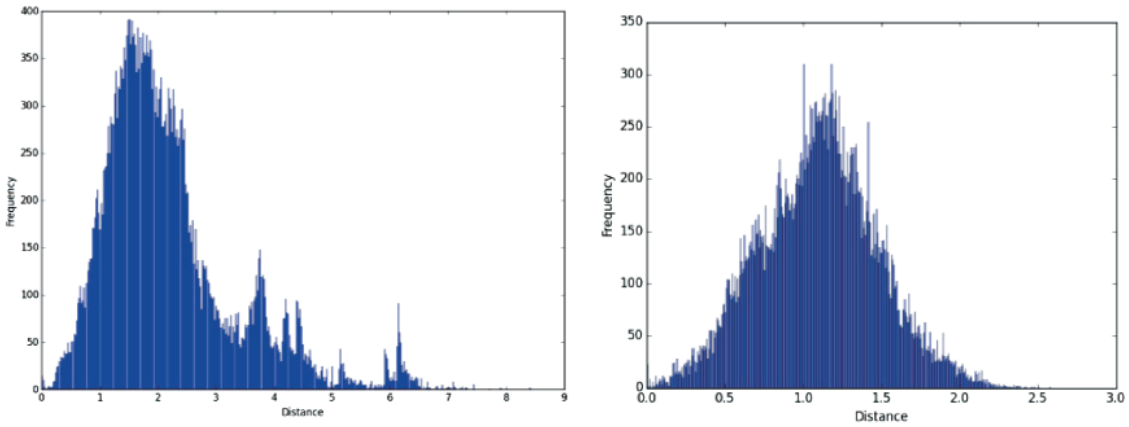

Рис. 2. Гистограммы расстояний между объектами: исходные данные без нормализации (слева), нормализованные исходные данные (справа)

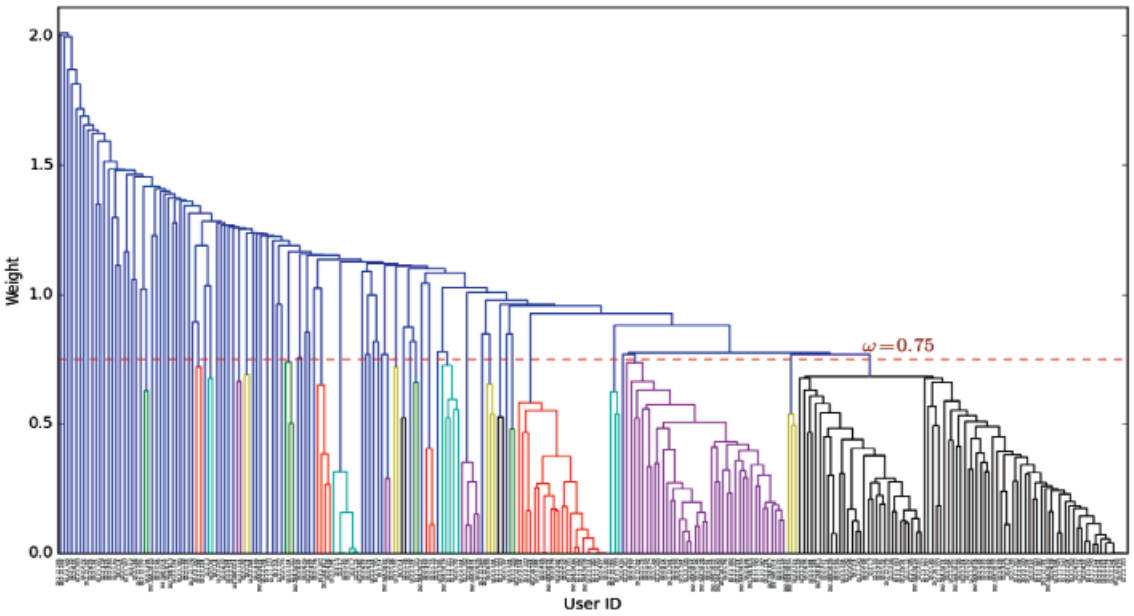

Рис. 3. Таксономия пользователей на основе анализа их тематических предпочтений. Иерархия составлена агломеративным усредняющим методом

Результаты работы алгоритма иерархической кластеризации удобно представлять графически в виде дендрограммы. При этом, как правило, задают некоторое пороговое значение уровня значимости для выделения отдельных кластеров. На рис. 3 представлена дендрограмма, изображающая кластеры, полученные при заданном пороговом значении, равном 0,75. Кластеры выделены различными цветами.

Из дендрограммы видно, что около половины объектов формируют четко выраженные кластеры. Остальная же часть объектов довольно разнородна. Это может свидетельствовать о том, что интересы и предпочтения данной части пользователей довольно индивидуальны (уникальны).

Имея представление о желаемом количестве кластеров, воспользуемся методом *k*-средних (*k*-means) [3] для проверки достоверности полученных ранее результатов и оценки их правдоподобности. Введем функционал качества кластеризации Φ_0, равный сумме средних внутрикластерных расстояний [3]:

$$\Phi_0 = \sum_{y \in Y} \frac{1}{|K_y|} \sum_{i: y_i = y} \rho^2(x_i, \mu_y) \to \min, \qquad (*)$$

где $K_y = \{x_i \in X^l \mid y_i = y\}$ – кластер с номером $y \in Y$, X^l – множество некластеризованных объектов; μ_y – центры кластеров y, ρ – евклидова метрика расстояния.

Экспериментируя с количеством кластеров и оценивая значения функционала качества (*), а также других статистических характеристик, таких как внутрикластерная дисперсия и стандартное отклонение, выберем наиболее подходящую модель кластеризации.

На рис. 4 представлены результаты кластеризации, полученные с помощью онлайн сервиса BigML Machine Learning Made Easy [8] для 16 кластеров.

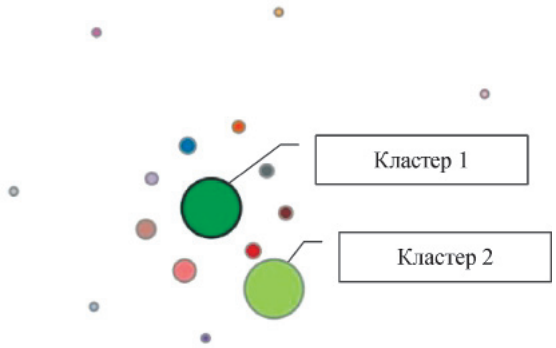

Рис. 4. Результаты работы алгоритма KMeans для 16 кластеров

Как видно из рис. 4, основная масса объектов (пользователей) разделилась на две большие группы (кластеры 1 и 2). Остальные объекты образуют мелкие кластеры. В целом такая картина подтверждает вышеописанные результаты, полученные с помощью алгоритма иерархической кластеризации.

Центры кластеров позволяют судить о признаках (рубриках), оказавших наибольшее влияние на их формирование. В частности, на формирование кластеров 1 и 2 (рис. 4) наибольшее влияние оказали рубрики «Интернет», «Социальные сети» и «Хостинги видео» соответственно. Чтобы детально оценить тематические предпочтения пользователей, входящих в эти кластеры, абстрагируемся от остальных признаков и представим объекты в виде трехмерной диаграммы, где каждая ось соответствует одному из признаков (рис. 5).

На рис. 5 видны три кластера. Пользователи кластера 1 преимущественно посещают Интернет-порталы, пользуются поисковыми системами. Пользователи кластера 2, напротив, уделяют большую часть времени просмотру видео и нахождению в социальных сетях. На предыдущем этапе с использованием метода k-means эти кластеры были объединены в один. Кластер 3 состоит из активных пользователей социальных сетей. Здесь вес остальных рубрик достаточно низкий, что вполне объяснимо, т.к. данная категория пользователей получает большую часть нужной им информации именно из социальных сетей, в том числе это касается и просмотра видео.

Говоря о нецелевом использовании ресурсов Интернет и эффективности использования рабочего времени сотрудниками, можно установить некоторый порог, поместив в начало координат сферу радиуса R. Объекты (пользователей), выходящие за границы этой сферы, будем считать расходующими ресурсы Интернет на нецелевые нужды.

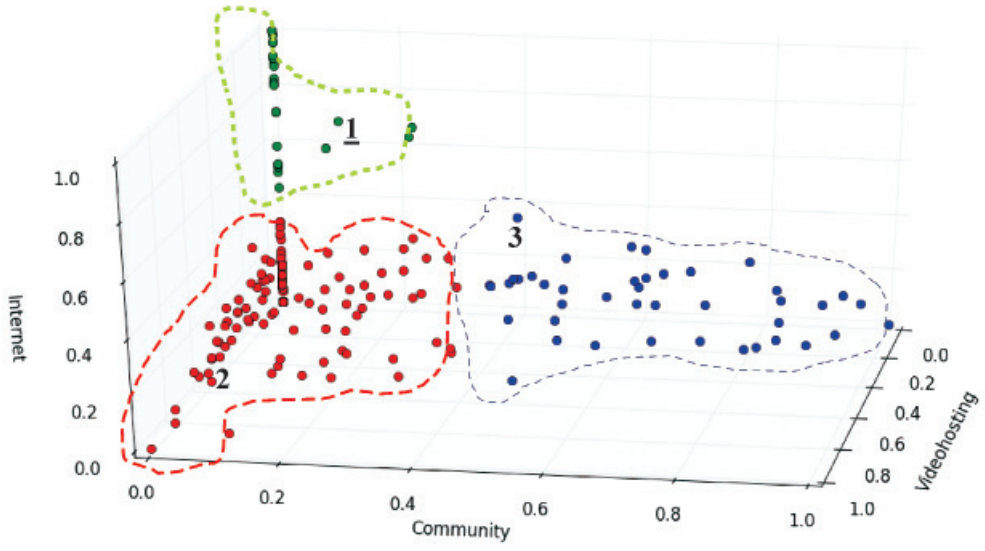

Рис. 5. Кластеры пользователей, соответствующие рубрикам «Интернет», «Соц. сети» и «Хостинги видео»

Результаты кластеризации, безусловно, зависят от многих факторов, в том числе от процесса подготовки исходных данных и выбранного метода кластеризации. Но в целом можно считать, что полученные кластеры позволяют достаточно адекватно определить предпочтения пользователей относительно выбора информационных ресурсов, просматриваемых в сети Интернет. Исследовав набор тематических категорий, оказавших наибольшее влияние на формирование кластеров, можно сделать вывод не только об информационных потребностях пользователей, но и о том, насколько эффективно сотрудники организации используют свое рабочее время.

Список литературы

1. Как узнать больше о ваших пользователях? Применение Data Mining в Рейтинге Mail.Ru [Электронный ресурс]. – Режим доступа: http://habrahabr.ru/company/mailru/blog/244285/ (дата обращения: 01.07.15).

2. Леонова Ю.В., Федотов А.М. Исследование пользовательских предпочтений для контроля и оптимизации Интернет-трафика в организации // Труды 11-й Всероссийской научной конференции «Электронные библиотеки: перспективные методы и технологии, электронные коллекции» – RCDL'2009. – Петрозаводск, Россия, 2009. – С. 158–166.

3. Машинное обучение (курс лекций, К.В. Воронцов) [Электронный ресурс]. – Режим доступа: http://www.machinelearning.ru (дата обращения: 01.07.15).

4. Паутов К.Г., Попов Ф.А. Информационная система анализа и тематической классификации веб-страниц на основе методов машинного обучения // Современные проблемы науки и образования. – 2012. – № 6. – С. 66.

5. Паутов К.Г., Попов Ф.А. Метод тематической классификации веб-страниц и его применение в системе учета и контроля Интернет-трафика вуза // Труды XVIII Всероссийской научно-методической конференции Телематика'2011 (20–23 июня 2011 года, Санкт-Петербург). – СПб: Санкт-Петербургский государственный университет информационных технологий, механики и оптики, 2011. – С. 21–22.

6. Паутов К.Г., Попов Ф.А. Обнаружение групп пользователей со схожими предпочтениями для контроля и оптимизации интернет-трафика в вузе // Труды XXI Всероссийской научно-методической конференции Телематика-2014 (23–26 июня 2014 года, Санкт-Петербург). – СПб.: Санкт-Петербургский национальный исследовательский университет информационных технологий, механики и оптики, 2014. – С. 178–179.

7. Фонд Общественное мнение [Электронный ресурс]. – Режим доступа: http://fom.ru (дата обращения: 01.07.15).

8. BigML Machine Learning Made Easy [Электронный ресурс]. – Режим доступа: http://bigml.com (дата обращения: 01.07.15).

References

1. Kak uznat bolshe o vashikh polzovatelyakh? Primenenie Data Mining v Reytinge Mail.Ru Available at: http://habrahabr.ru/company/mailru/blog/244285/ (accessed: 01.07.15).

2. Leonova Yu., Fedotov A. Issledovanie polzovatelskikh predpochteniy dlya kontrolya i optimizatsii Internet-trafika v organizatsii [Research of the user preferences for the control and Internet traffic optimisation in the organization]. Proc. of the *XI All-Russian Research Conference RCDL2009 «Digital Libraries: Advanced Methods and Technologies, Digital Collections»*. Petrozavodsk, September 17–21, 2009, pp. 158–166.

3. Mashinnoe obuchenie (kurs lektsiy, K.V. Vorontsov) Available at: http://www.machinelearning.ru (accessed: 01.07.15).

4. Pautov K.G., Popov F.A. Informatsionnaya sistema analiza i tematicheskoy klassifikatsii veb-stranits na osnove metodov mashinnogo obucheniya [An information system of web-page classification based on machine learning methods], *Sovremennye problemy nauki i obrazovaniya,* 2012, No. 6, p. 66, Available at: http://www.science-education.ru/106-7680 (accessed: 01.07.15).

5. Pautov K.G., Popov F.A. Metod tematicheskoy klassifikatsii veb-stranits i ego primenenie v sisteme ucheta i kontrolya Internet-trafika vuza. *Trudy XXI Vserossiyskoy nauchno-metodicheskoy konferentsii Telematika2011.* (Saint Petersburg, 20–23 June 2011). Saint Petersburg National Research University of Information Technologies, Mechanics and Optics. Saint Petersburg, 2011, pp. 21–22.

6. Pautov K.G., Popov F.A. Obnaruzhenie grupp polzovateley so skhozhimi predpochteniyami dlya kontrolya i optimizatsii internet-trafika v vuze. *Trudy XXI Vserossiyskoy nauchno-metodicheskoy konferentsii Telematika2014.* (Saint Petersburg, 23–26 July 2014). Saint Petersburg National Research University of Information Technologies, Mechanics and Optics. Saint Petersburg, 2014, pp. 178–179.

7. Fond Obshchestvennoe mnenie Available at: http://fom.ru (accessed: 01.07.15).

8. BigML Machine Learning Made Easy Available at: http://bigml.com (accessed: 01.07.15).

Рецензенты:

Старовиков М.И., д.п.н., к.ф.-м.н., доцент, профессор кафедры физики и информатики, ФГБОУ ВПО «Алтайская государственная академия образования им. В.М. Шукшина», г. Бийск;

Попок Н.И., д.т.н., профессор, нач. лаборатории ОА «ФНПЦ «Алтай», г. Бийск.