Группировка пользователей Интернета, основанная на истории их веб-сессий

© Юлия Киселева

Научный руководитель: Борис Асенович Новиков

Санкт-Петербургский государственный университет
julianakiseleva@gmail.com

Аннотация

В данной работе рассматривается вопрос персонализации пользователей Интернета на основе групп, отражающих интересы. В настоящее время является наиболее важным для исследований в области логического анализа данных Интернета. Существующие подходы группировки веб-пользователей основываются на снэпшотах их веб-сессий. Данный подход описан в [1]. Группы пользователей Интернета образуются, исходя из истории их веб-сессий, широко используется в области веб-рекламы

1 Введение

Логический анализ данных Интернета – набор подходов для выявления шаблонов поведения пользователей. На данный момент является активной областью для исследований [1]. Существующие подходы и техники включают в себя статистический анализ [1], последовательные шаблоны [2], ассоциативные правила [3], классификацию [4] и другие методы. Важным аспектом логического анализа данных Интернета является выявление групп пользователей с близкими интересами.

Существующие подходы для группировки веб-пользователей состоят из трех этапов:

1. Подготовка данных - данная фаза представляет собой извлечение данных с сервера, затем проводится идентификация пользователей и их веб-сессий.
2. Выявление групп пользователей.
3. Анализ полученных групп.

Веб-данные по природе своей являются динамически развивающимися, и как следствие этого факта существуют два подхода для группировки пользователей:

1. возможность выявления похожих пользователей в процессе эволюции данных;
2. возможность выявления похожих пользователей за фиксированный промежуток времени. В этом случае результат группировки нуждается в постоянном обновлении, по мере поступления новых данных.

В данной статье мы сфокусировались на втором подходе выявления похожих пользователей. Проанализировали результаты – группы веб-пользователей, которые были полученные с использованием двух метрик, введенных в работе для измерения близости между пользователями. Целью представленной работы является получение единой методики оценки близости между пользователями Интернета.

2 Метрики похожести пользователей

2.1 Набор данных

Для исследования был использован лог, который содержит информацию о 1343 пользователях, общее количество запросов в нем – 66380. Будем называть запросы одного пользователя документом. Далее для построения групп схожих пользователей будем использовать поисковые их запросы.

2.2 «Очистка» данных

Перед началом эксперимента следует очистить данные, для этих целей можно использовать WordNet. WordNet – это большая лексическая база данных английского языка [6], при помощи него можно избавится от опечаток, совершенных пользователями. Так же помощью отдельного фильтра убираем стоп-слова, такие как «how», «and» и другие. Проведенная очистка уменьшила рассматриваемое множество на 1.4%.

2.3 Обработка данных

Целью исследования является нахождение наиболее близких пользователей и объединение их в группы. Предполагаем, что пользователи, попавшие в одну и ту же группу, обладают схожими интересами. В данной работе используем две метрики для определения близости между запросами пользователей, они описаны ниже. Сначала создаем пространство всех слов, которые встретились в доку-

Труды 10-й Всероссийской научной конференции «Электронные библиотеки: перспективные методы и технологии, электронные коллекции» – RCDL'2008, Дубна, Россия, 2008.

ментах. Затем для каждого пользователя получаем вектор весов слов, которые встречаются в его и только в его документе (в наборе его запросов). Вектор выглядит следующим образом:

$$d_j = w(t_1), w(t_2) \ldots w(t_n), \quad (1)$$

где $w(t_j)$ – это вес *tf–idf* слова t_j во всем множестве слов (term frequency–inverse document frequency) [5], где *tf* частота слова t_j:

$$\mathrm{tf_{i,j}} = \frac{n_{i,j}}{\sum_k n_{k,j}}$$

в представленном выражении n_{ij} – это сколько раз встретилось рассматриваемое слово в документе d_j, а знаменатель представляет собой количество всех слов в документе d_j.

Обратная частота – это мера, показывающая общую важность слова:

$$\mathrm{idf_i} = \log \frac{|D|}{|\{d_j : t_i \in d_j\}|}$$

- где | *D* | общее количество документов в наборе.
- $|\{d_j : t_i \in d_j\}|$ количество документов, в которых встречается слово t_i.

Согласно второй метрике, каждый пользователь представлен набором векторов, координаты которых представляют собой веса слов, встретившихся в его запросе. Вектор-запрос представлен в следующем виде:

$$Q_i = w(t_1), w(t_2) \ldots w(t_n), \quad (2)$$

где $w(t_j)$ – это вес слова *j* в запросе Q_i.

2.4 Метрика (1): средняя мера близости

Итак, выше определи метрики, которые помогут нам объединить пользователей с похожими интересами в группы.

Первая метрика, представленная в 2.2, для определения близости между пользователями использует скалярное произведение векторов их весов. Ясно, что наиболее близкие пользователи имеют наибольшее значение скалярного произведения их весовых векторов.

В результате проведения эксперимента получаем матрицу близости между всеми пользователями, каждый элемент матрицы $\{a_{ij}\}$ это мера близости между *i*-м и *j*-м пользователями. Соответственно, $a_{ij} \in [0,1]$.

2.5 Метрика (2): максимальная схожесть запросов пользователей

Вторая метрика, определенная в 2.2, определяет близость между пользователями как максимум скалярного произведения весовых векторов их запросов. Соответственно, если пользователи имеют два одинаковых запроса, то они имеют близость = 1. Такая метрика не чувствительна в случаях, когда пользователи U_1, U_2 и U_3 имеют одинаковые запросы (например, “vegas”), близость между этими пользователями равна 1, но мы можем встретить ситуацию, когда U_1 ввел запрос “vegas” 10 раз, U_2 – 8 раз, а U_3 – только 1 раз. В подобных ситуациях определенная выше метрика работает не слишком удачно.

2.6 Диаграммы распределения, полученных метрик

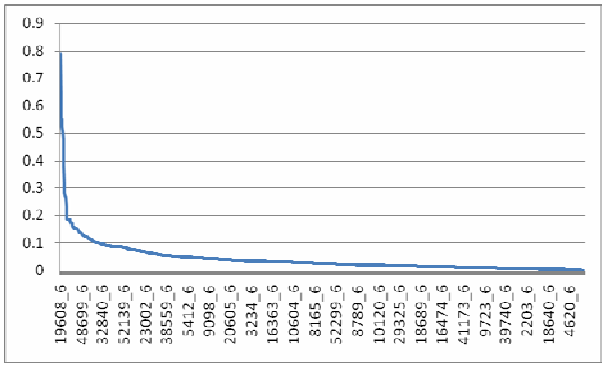

Диаграмма 1: распределение величин, полученное на основе первой метрики, ось x – user_id, ось y – величины первой метрики.

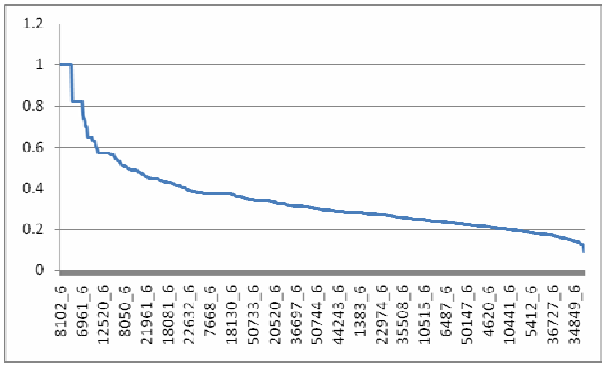

Диаграмма 2: распределение величин, полученное на основе второй метрики, ось x – user_id, ось y – величины второй метрики.

3 Полученные результаты и будущие эксперименты

Основной целью исследования является создание метрики для нахождения схожих пользователей, которая будет лишено недостатков обнаруженных, в процессе экспериментов, в представленных метриках. Для оценки полученных данных были использованы метод «Общего котла» [7] и анализ тематических срезов данных, т.е. рассматривался срез данных, который содержал одно или несколько тематических слов. Затем асессорами оценивалось, насколько реально близки запросы, находящиеся в полученном срезе.

Итогом исследований представляется создание карты интересов пользователей Интернета, что является важным особенно при показе тематической рекламы и также упорядоченности информации, извлечение полезных знаний.

Литература

[1] C. Buchwalter, M.Ryan, and D. Martin. The state of online advertising: data covering 4th Q 2000. In TR Adrelevance, 2001.

[2] Q. Yang, H.H. Zhang, and T. Li. Mining web logs for prediction models in www caching and prefetching. In *Proc.of ICCNMC'01*, 2001.

[3] B. Mobasher, H.Dai, T. Luo, and M. Nakagawa. Effective personalization based on association rule discovery from the web usage data. In Proc. Of WIDM, 2001.

[4] T. Li, Q. Yang,and K. Wang. Classification pruning for web-request prediction. In Proc.of WWW, 2001.

[5] Baeza-Yates, R., Hurtado, C., Mendoza, M.: Query recommendation using query logs in search engines. In: Current Trends in Database Technology – EDBT, Springer-Verlag GmbH (2004) 588–596

[6] Agirre, Eneko and David Martinez. "Integrating selectional preferences in WordNet." In: Proceedings of the first International WordNet Conference, Mysore, India, 21-25 January 2002.

[7] И. Некрестьянов, М. Некрестьянова, А. Нозик. К вопросу об эффективности метода «общего котла» //Труды RCDL'2005. – Ярославль, 2005.

Grouping Web-users based on Query Log

Julia Kiseleva

Web pages are personalized based on the interests of an individual. Personalization implies that the changes are based on implicit data, such as items purchased or pages viewed. In our research we don't approach to strongly user's personalization. Generally, typical web user grouping approach consists of three phases: data preparation, group discovery and group analysis. This is work in progress report. At this stage of our research we focus on user similarity metrics that later will be user to group users. In this report we present description of our approach, define several metrics and conduct experiments to evaluate their quality.