УДК 004.912

М.В. Бочков, П.Н. Бойков

ОПТИМИЗАЦИЯ СТРАТЕГИИ ПОИСКА ПОЛЬЗОВАТЕЛЯ В СОЦИАЛЬНЫХ СЕТЯХ

Бочков Максим Вадимович, *д.т.н., окончил факультет многоканальной электросвязи Орловского высшего военного командного училища связи. Профессор НОУ ДПО «Центр предпринимательских рисков», Санкт-Петербург.[e-mail: mvboch@yandex.ru].*

Бойков Павел Николаевич, *окончил факультет информационно-телекоммуникационных систем Академии Федерального агентства правительственной связи и информации при Президенте РФ. Ведущий специалист ОАО «НИИ «Рубин», Санкт-Петербург. [e-mail: boykovpn@yandex.ru].*

Аннотация

В статье исследованы закономерности представления пользователями социальных сетей своих данных, на основе которых разработан алгоритм формирования оптимальной стратегии поиска в социальных сетях.

Ключевые слова: социальная сеть, поисковая функция, поисковая оптимизация, дерево решений.

Социальная сеть как объект исследования

Среди ресурсов в сети Интернет все большую популярность приобретают онлайновые социальные сети (ОСС). К типовым возможностям их участников можно отнести:

– обмен информационными ресурсами с другими участниками ОСС;
– публикация и обсуждение идей;
– выбор социальных групп (сообществ) и участие в них;
– использование развлекательных и досуговых сервисов ОСС и др.

Очевидной тенденцией в развитии ОСС является рост числа пользователей и развитие их функциональных сервисов [1]. Динамика изменения числа пользователей в наиболее популярных ОСС представлена на рисунке 1.

Информационную основу ОСС образуют персональные пользовательские страницы. Как правило, создатели ОСС стремятся получить от пользователя максимум информации. С этой целью регистрационная форма предлагает опубликовать максимум идентификационной и другой персональной информации. На рисунке 2 показан набор регистрационных данных в ОСС «Вконтакте».

Очевидно, что наиболее полное представление пользователями своих данных повышает точность и полноту результатов запроса, а следовательно, однозначность идентификации участников ОСС. С другой стороны, среднестатистический пользователь подсознательно стремится представить минимум информации

о себе, ограничить круг своего общения, обеспечив себе комфортное общение в ОСС. Таким образом, проявляется конфликт интересов владельцев и пользователей ОСС – одни хотят знать все, а другие хотят обойтись минимум информацией о себе [2].

Целью настоящего исследования является выявление закономерностей представления идентифицирующей пользователя информации в ОСС. Знание таких закономерностей и их описание в виде формальной модели позволит сформировать оптимальную стратегию поиска, при которой вероятность точного нахождения требуемого пользователя ОСС за минимальное число итераций поиска будет максимальна.

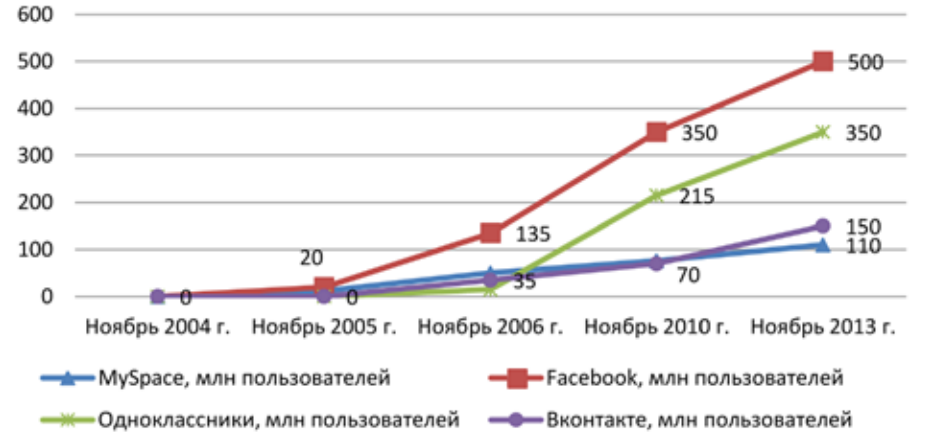

Рис. 1. Динамика изменения числа пользователей в ОСС

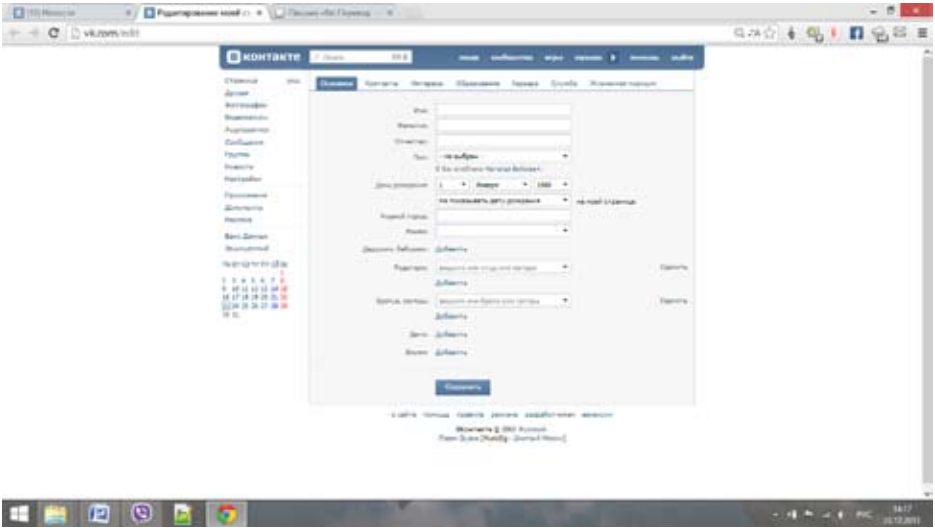

Рис. 2. Набор регистрационных данных в ОСС «Вконтакте»

Модель представления информации пользователями ОСС

Исходные данные

В настоящем исследовании использован модельный фрагмент ОСС, сформированный путем обезличивания репрезентативного дампа общедоступных в сети Интернет пользовательских страниц. На основе полученной информации проведен расчет статистик, характеризующих атрибуты регистрационных данных пользователей [3].

Ранжирование пользовательских атрибутов и интерпретация полученных результатов

Для последующих исследований были выделены следующие пользовательские атрибуты:

– идентификатор пользователя в ОСС (a_0);

– фамилия (a_1);

– имя (a_2);

– город проживания (a_3);

– пол пользователя (a_4);

– дата рождения (a_5);

– наименование и год окончания вуза (a_6);

– наименование и год окончания школы (a_7);

– место работы (a_8);

– семейное положение (a_9).

Для исследований были введены следующие упрощения:

1. Параметр a_0 не использован при анализе, так как является уникальным для каждого пользователя и однозначно идентифицирует его в социальной сети.

2. Параметры a_4 и a_9 исключены из анализа ввиду малого диапазона принимаемых значений и незначительного влияния на результаты поиска.

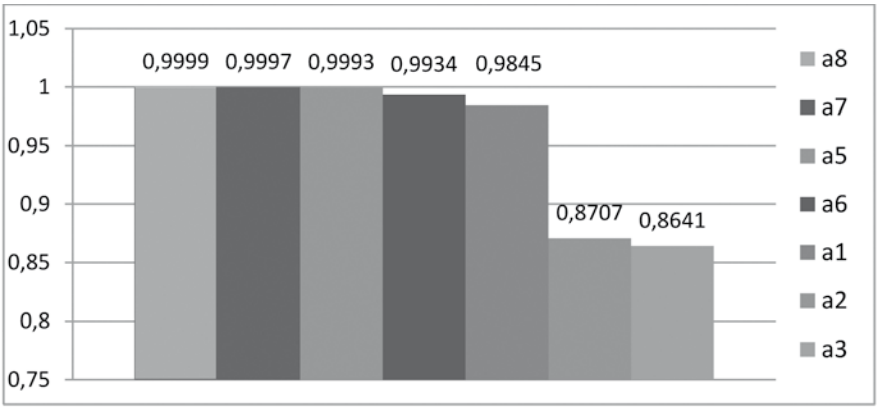

Рис. 3. Ранжирование вероятностей встречаемости атрибутов на пользовательских страницах ОСС

3. Значения атрибута a_2 аналогичные: «Иван», «Ваня», «Ванька» считаются эквивалентными поисковому запросу «Иван».

4. Для атрибута a_3 значения, подобные «СПб», «Санкт-Петербург», «Питер», считались эквивалентными поисковому запросу «Санкт-Петербург».

На рисунке 3 представлены ранжированные значения вероятностей присутствия пользовательских атрибутов, отражающих закономерности отображения информации в ОСС. В дальнейшем данная закономерность используется в качестве модели представления информации пользователями ОСС.

Алгоритм формирования максимальной стратегии

Использование модели представления, в свою очередь, позволяет на основе математического аппарата инфодинамического моделирования [4] решить задачу формирования оптимального алгоритма поиска пользователя. В соответствии с ним для отражения «привлекательности» поискового атрибута выбраны показатели условной и взаимной энтропии.

1. Условная энтропия $H(f|\alpha_j)$ функции поиска f при заданном значении атрибута α_j:

$$H\left(f|a_j\right)=\sum_{\beta=0}^{1} p\left(\beta a_j\right) I\left(\beta \mid a_j\right),$$

где β – значение поисковой функции (0 – результат поиска не удачен, 1 – поиск успешен).

2. Взаимная информация $I(f;a_j)$ между поисковой функцией f и атрибутом a_j определяется выражением

$$I\left(f;a_j\right)=\sum_{\beta=0}^{1} p\left(\beta a_j\right) I\left(\beta; a_j\right).$$

Результаты расчетов значений условной энтропии поисковой функции и взаимной информации между поисковой функцией и значением каждого атрибута a_j показаны в таблице 1.

Таблица 1

Результаты расчетов

	a_1	a_2	a_5	a_6	a_7	a_8
$H(f\|\alpha_j)$	0,358	0,158	0,363	0,358	0,396	0,279
$I(f;\alpha_j)$	0,802	0,698	0,681	0,802	0, 936	1,049

Для выработки логически обоснованных критериев выбора состава и порядка атрибута при разработке стратегии поиска необходимо более детально проанализировать содержательную сторону приведенных формальных оценок. Для этого представим связь атрибута и поисковой функции следующим образом:

$$H(f)-H(f|a_j)=I(f;a_j)=H(a_j)-H(a_j|f).$$

Рассмотрим содержательно слагаемые этого выражения.

1. Энтропия $H(f)$ поисковой функции – среднее количество информации, которое необходимо извлечь для определения значения функции.

2. Энтропия $H(a_j)$ атрибута – среднее количество информации, которое извлекается при добавлении атрибута поиска.

3. Взаимная информация $I(f;a_j)$ – среднее количество информации о результатах поиска, которое несет атрибут поиска.

Для решения поставленной задачи интерес представляет взаимная информация, как индикатор того, насколько уменьшится диапазон результатов поиска при наличии того или иного атрибута. Другими словами, из информации $H(a_j)$ оценивается та ее часть, которая позволяет уменьшить энтропию $H(f)$ функции до значения $H(f|\alpha_j)$.

Под стоимостью решения будем понимать время отклика на выполнение поискового запроса и примем его одинаковым для каждого атрибута поиска. Таким образом, критерием оптимизации при выборе следующего теста будет выступать выражение

$$a_j^* = \max_{a_j \in A} I(f, a_j),$$

где a_j^* – следующий атрибут для добавления в запрос.

В общем виде задача формирования оптимальной стратегии поиска в терминах инфодинамического моделирования соответствует задаче конструирования деревьев решений. Процедуру конструирования дерева решений представим в виде алгоритма.

Шаг 1. Задача состоит в выборе атрибута, который целесообразно использовать первым. Результаты вычислений (табл. 2) показывают, что критерию оптимизации удовлетворяет атрибут a_8. Построим первый уровень дерева решений при условии, что корневому узлу дерева соответствует этот атрибут. Как показывает таблица решений, выбор данного атрибута не позволит идентифицировать пользователя, поэтому на основе вероятности совместной встречаемости атрибутов целесообразно в качестве второго атрибута использовать a_1. Если атрибут a_1 отсутствует, то в качестве второго атрибута выбирается атрибут, следующий по значению совместной вероятности встречаемости атрибутов с a_8.

Шаг 2. Если атрибут a_8 отсутствует, то выбирается следующий по значимости атрибут, а алгоритм добавления второго атрибута в параметры запроса аналогичен шагу 1.

Шаг 3. На этом шаге выбирается переменная $a_8 = a_7 = 0$. Из таблицы решений видно, что поисковая функция f принимает максимальное значение только в том случае, если будут известны следующие пары атрибутов: α_1 α_5 и α_5 α_6. Таким образом, целесообразно проверить общий для этих пар атрибут α_5. Если он от-

сутствует, то поиск целесообразно прекратить, так как оставшихся атрибутов недостаточно для идентификации пользователя в социальной сети.

Вычисление информационных оценок (для таблицы 1) и выбор переменных при построении дерева решений приведены в таблице 2.

Таблица 2

Вычисление информационных оценок и выбор переменных

Уровень дерева решений	Условие (известные значения переменных)	Переменные, из которых осуществляется выбор	Взаимная информация $I(f;a_j)$	Выбор
1	–	α_1 α_2 α_5 α_6 α_7 α_8	0,802 0,698 0,681 0,802 0,936 **1,049**	α_8
2	$\alpha_8 = 1$	α_1 α_2 α_5 α_6 α_7	0,802 0,698 0,681 0,802 **0,936**	α_7
3	$\alpha_7 = 1$	α_1 α_2 α_5 α_6	0,802 0,698 0,681 **0,802**	α_1
4	$\alpha_1 = 1$	α_2 α_5 α_6	0,698 0,681 **0,802**	α_6
5	$\alpha_6 = 1$	α_2 α_5	**0,698** 0,681	α_2

Дерево решений как результат оптимизации представлено на рисунке 4.

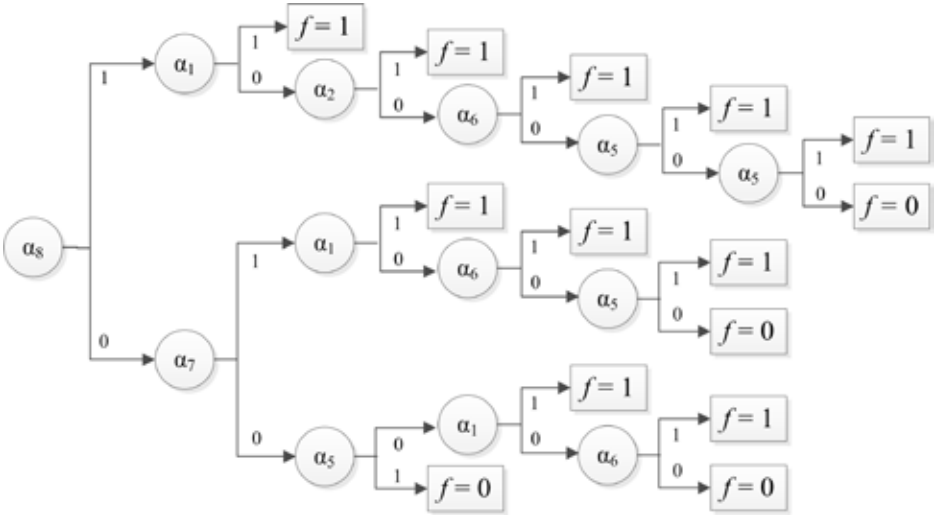

Рис. 4. Дерево решений на основе информационных оценок

Экспериментальные исследования полученного алгоритма поиска пользователя социальной сети

На заключительном этапе исследовалась практическая пригодность полученного алгоритма для поиска пользователя в социальной сети, в частности, определялось, возможно ли конструирование множества деревьев решений за приемлемое время. Затем оценивалось преимущество предложенного алгоритма по сравнению с алгоритмом последовательного добавления атрибутов поиска в поисковый запрос.

Для получения априорных сведений о пользователях социальной сети рассматривались три наиболее популярные социальные сети: «Вконтакте», «Facebook» и «Одноклассники». Выборка осуществлялась путем случайного отображения информации о пользователе социальной сети (такая функциональная возможность предоставляется ОСС). В результате эксперимент проводился на общей выборке из 4500 случайно выбранных пользователей социальных сетей.

В ходе эксперимента были получены следующие результаты. Во всех трех социальных сетях алгоритм поиска пользователя сконструировал деревья, имеющие меньшую стоимость по сравнению с деревьями решений алгоритма последовательного добавления атрибутов поиска в поисковый запрос. В среднем стоимость деревьев решений на 8 % меньше (рис. 5). Также снизились и вычислительные затраты. Так, в среднем для успешного поиска пользователя в социальных сетях необходимо использовать три атрибута, а не четыре.

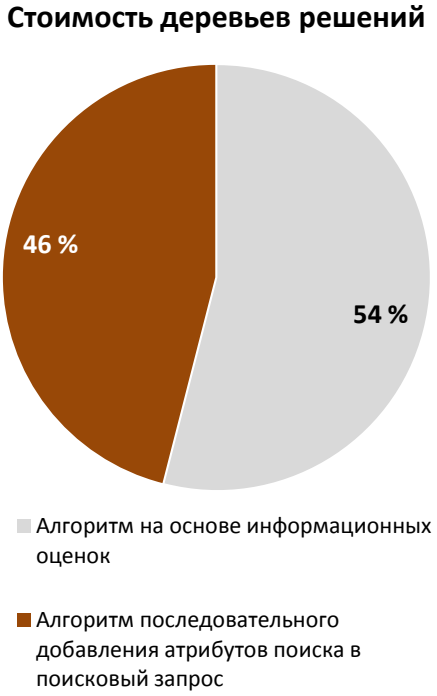

Рис. 5. Сравнение вычислительных затрат двух алгоритмов для успешного поиска пользователя в социальной сети

Выводы

1. Предложенный алгоритм конструирования дерева решений на основе информационных оценок пользовательских атрибутов позволяет оптимизировать время поиска пользователей в ОСС при достаточно большом числе атрибутов поиска.

2. Использование разработанного алгоритма позволяет сократить вычислительные затраты на обработку сервером социальной сети поискового запроса, тем самым способствуя сокращению времени, затраченного аналитиком на поиск путем последовательного добавления пользовательских атрибутов в поисковый запрос.

3. Полученный подход применим к любым базам данных, содержащим большое число объектов учета с множеством идентифицирующих и характеризующих объекты атрибутов, что позволяет говорить о формировании универсальной стратегии поисковой оптимизации.

СПИСОК ЛИТЕРАТУРЫ

1. Губанов Д.А., Новиков Д.А., Чхартишвили А.Г. Социальные сети: модели информационного влияния, управления и противоборства. – М. : МЦНМО, 2010.

2. Бочков М.В., Бойков П.Н., Яшин А.А. Социальные сети как основной источник утечки персональных данных // Inside. Защита информации. – 2010. – № 3.

3. Бочков М.В., Бойков П.Н. Способ автоматического рубрицирования неструктурированной информации в сети Интернет // Информационные технологии. – 2012. – № 2.

4. Курбацкий А.Н., Чеушев В.А. Информационный метод анализа и оптимизации в системах поддержки принятия решений. – Минск: Ин-т техн. кибернетики НАН Беларуси, 1999.