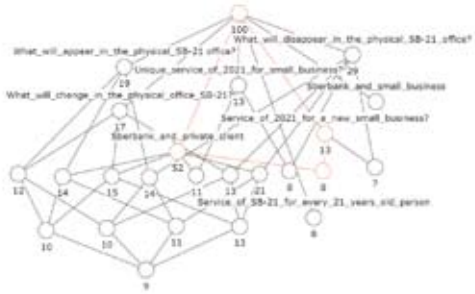
Анализ данных в краудсорсинговых проектах



Число успешных краудсорсинговых проектов растет, сообщества их экспертов генерируют все больше данных, и для понимания происходящих в коллаборативных платформах процессов требуются новые методы анализа.

Ключевые слова: коллаборативные технологии, анализ данных, скрытые сообщества, анализ активности пользователей, социальные сети

Дмитрий Игнатов,
Александра Каминская,
Андрей Константинов

Успехи современной индустрии коллаборативных технологий ознаменовались появлением ряда новых платформ для проведения распределенных «мозговых штурмов» или общественной экспертизы. Среди них, например, решения таких компаний, как Spigit, Jive, BrightIdea, InnoCentive и Imaginatik, а в 2011 году действующие краудсорсинговые платформы появились и в России — Witology и Wikivote. И хотя до технологического прорыва еще далеко, уже имеется несколько крупных краудсорсинговых проектов («Сбербанк-21», «Национальная предпринимательская инициатива» совместно с АСИ и др.).

Как правило, в рамках одного проекта решается общая задача, выдвигаются и оцениваются идеи, а в итоге, по результатам обсуждений и рейтингования, определяются лучшие предложения и их генераторы. Ядро таких краудсорсинговых систем составляют социосемантические сети, которые, в отличие от социальных, кроме пользователей описывают дополнительные объекты (в частности, выдвигаемые идеи) и отношения между ними и пользователями, семантически наполняя сети. В современных социальных сетях такое семантическое наполнение тоже присутствует — в специализированных группах здесь обсуждаются темы, фотографии, посты пользователей, однако все это не рассматривается как значимые сущности, а является лишь средством для общения пользователей (одним из типов связей). Что же касается социосемантических сетей и краудсорсинговых проектов, проводимых на их основе, то все предложения пользователей по решению некоторой задачи могут стать значимыми идеями, а обсуждения и оценки таких предложений другими пользователями отсеивают слабые и отбирают лучшие.

Для глубокого понимания поведения пользователей, выработки адекватных критериев оценки их работы, анализа динамики оценивания и получения прочей статистики в ходе развития проекта необходимы особые средства. Традиционные методы кластеризации, поиска сообществ и анализа текстов нуждаются в адаптации, а иногда и в значительной переработке. Поэтому, например, для того чтобы устранить потерю общего описания объектов кластеров, предлагается бикластеризация, а для учета триадического характера данных <объект, признак, условие> или <пользователь, ключевое слово, идея> применяются оригинальные методы трикластеризации, реализованные, в частности, в системе анализа данных коллаборативных платформ CrowDM (Crowd Data Mining). Для описания моделей анализа данных краудсорсингового проекта в данной системе применяется **анализ формальных понятий (АФП)** — алгебраический подход к анализу данных, предназначенный для исследования объектно-признаковых данных. Такие данные удобно представлять в виде таблиц, в которых строкам соответствуют объекты, а столбцам — признаки. Таблица с указанием того, каким признаком обладают ее объекты, называется формальным контекстом.

Объектно-признаковая группа, в которой каждый объект обладает всеми признаками, присущими данной группе объектов, называется формальным понятием (в таблице формальному понятию соответствует максимальный по вложению полностью заполненный крестиками прямоугольник). Таким образом, формальное понятие состоит из множества объектов (объема понятия), каждый из которых обладает некоторым множеством признаков (содержание понятия), с тем условием что больше никакой объект в таблице всеми этими признаками не обладает (аналогично для признаков). Далее можно изучать найденные формальные понятия, выискивая сгруппированные вместе интересные группы объектов и признаков, или построить иерархию найденных понятий по вложению их объемов — решетку понятий, что значительно упрощает навигацию.

Основная цель краудсорсинговых проектов состоит в выявлении толковых идей и их «генераторов» — людей. В платформе Witology эту деятельность направляют фасилитаторы — специально обученные люди. Если обычные модераторы форумов лишь наблюдают за дискуссией со стороны, изредка блокируя пользователей за нецензурную лексику или несоответствие теме, то фасилитаторы, кроме этого, выполняют и другие функции: задают темы для обсуждений (следуя требованиям заказчика), награждают лучших участников проекта («Герой недели»), напоминают пользователям о важных событиях проекта («Сегодня началось голосование за идеи!») и т. п., поддерживая активность краудсорсеров. Поэтому и поиском людей через идеи занимаются также фасилитаторы во главе с менеджером проекта. Таким образом, краудсорсинг — это направляемая деятельность творческих,

критично настроенных людей, решающих общую задачу.

А что, если выявлять не только толковых людей, но и группы пользователей по интересам, которые в дальнейшем могут перерасти в экспертные сообщества? Потом можно пойти дальше и искать зависимости в действиях краудсорсеров (например, «если пользователь предложил идею по теме 1, то он обязательно предложил идею и в теме 2»), а также давать рекомендации вновь прибывшим пользователям по поводу тем, обсуждение которых их могло бы заинтересовать. Таким образом, в краудсорсинге оказываются востребованы идеи анализа социальных сетей и рекомендательных систем. Отдельно стоит выделить семантический анализ контента, порожденного пользователем: выявление ключевых слов (у одного краудсорсера, у их группы, у всего проекта целиком, в обсуждении отдельных задач и идей); обнаружение полезного применения — например, признаковых зависимостей («если пользователь говорил о том, он поговорит и о другом»); обнаружение сообществ экспертов не в конкретной теме проекта, а экспертов по какому-то понятию; поиск слов-синонимов в контексте всего проекта и т. д. Здесь можно применять обычные статистические методы для анализа поведения пользователей — допустим, посмотреть на распределение количества выставленных оценок и оставленных комментариев и не только обогатить ежедневную статистику по проекту интересными выводами, но и, например, построить типологию пользователей краудсорсинговых систем.

Для разработки математических моделей, формализации и решения перечисленных задач компания Witology и НИУ ВШЭ создали совместную проектно-учебную группу «Алгоритмы интеллектуального анализа данных для интернет-форумов обсуждения инновационных проектов», результатом деятельности которой стало:

- создание моделей данных краудсорсинговых платформ на основе АФП, аналогичных по возможностям представления социо-семантическим сетям;
- разработка методов поиска тематических сообществ на основе бикластеризации и трикластеризации;
- создание модели рекомендательной системы единомышленников и «антагонистов», а также интересных идей для поддержания активности на платформе;
- анализ характера активности участников проекта «Сбербанк-21» и построение типологии пользователей краудсорсинговой платформы на его основе;
- исследование связи между рейтингами и стандартными характеристиками социальной сети (степени входящих и выходящих ребер узлов сети, меры центральности, «промежуточности» и т. п.);
- создание прототипа системы анализа краудсорсинговых платформ CrowDM.

ДАННЫЕ

Любой краудсорсинговый проект начинается с краткого описания постановок задач и предметной области — не стал исключением и проект «Сбербанк-21» по сбору мнений привлеченных экспертов о дизайне и функционале офиса Сбербанка 2021 года.

Для каждого типа офиса рассматривались по три проблемы — например, категория «Офис СБ-21 для индивидуальных предпринимателей» включала проблемы «Сбербанк и предприниматель: интерфейс 2021 года», «Услуга 2021 года для начинающих свое дело», «Уникальная услуга 2012 для предпринимателей». В ходе обсуждения идей требовалось, чтобы их описание включало формы и содержание решения, а также мотивацию и причины появления такого предложения. Предложения, прежде чем стать решениями, проходили через стадии эволюционного отбора и улучшения, включая: отбор схожих решений, общее голосование, доработку, биржу решений и рецензирование.

В предварительном тестировании участвовало более 5 тыс. человек, из которых 450 были отобраны и приглашены в качестве экспертов (33% — женщины, 67% — мужчины; 21% — сотрудники Сбербанка, 79% — клиенты и заинтересованные лица). В итоге 222 эксперта предложили 1581 решение по 15 задачам, сгруппированным по пяти темам. В ходе этапа «Отбор схожих решений» осталось 589 решений. За время работы эксперты провели на площадке в общей сложности порядка 4 тыс. человеко-часов, оставили 12 820 комментариев и выставили 66 896 оценок. В итоге среди предложенных решений три признаны лучшими.

МОДЕЛИ, АЛГОРИТМЫ И СХЕМА АНАЛИЗА

На начальном этапе анализа были выявлены два основных типа данных: без использования текстов, сгенерированных пользователями (связи, оценки, действия пользователей), и данные с текстами (ключевыми словами). И хотя социо-семантические сети не содержат явного разграничения на данные, использующие или нет ключевые слова, последние, скорее, относятся к социальной составляющей сети, а первые — к семантической. Поэтому для анализа данных без ключевых слов предлагается применять методы анализа социальных сетей (Social Network Analysis): поиск сообществ и ключевых пользователей; выделение компонентов связности; расчет мер центральности, влияния, «промежуточности» и близости и т. п. Текстовые данные второго типа требуют предварительной обработки перед этапом анализа.

Извлечение релевантных ключевых слов и словосочетаний — непростая задача, решаемая различными лингвистическими методами с применением статистических мер качества слов и привлечением экспертов предметной области. Деление данных можно провести еще на основе типов значений и характера их структуры. Так, данные отношения «человек А комментировал идею человека В» удобно представить в виде таблицы «человек-человек» с крестиками на пересечении строки А и столбца В или количеством комментариев. В таблицу «человек-идея» можно свести данные отношения «человек А предложил идею С». Это примеры двумерных (таблиц) бинарных и многозначных данных. Помимо двумерных, существуют и триадические данные, например, для отношения «человек А использовал ключевое слово В при комментировании идеи С». Подобное многообразие типов требует применения широкого арсенала методов: кластеризации, бикластеризации и трикластеризации, спектральной графовой кластеризации, анализа формальных понятий (решетки понятий, импликации, ассоциативные правила) и его расширений для случая мультимодальных данных, например триадических. Почему для выявления групп пользователей по интересам недостаточно только традиционных методов кластеризации? Дело в том, что стандартные подходы, такие как иерархическая кластеризация или метод k-средних, способны выявить группы схожих пользователей, но при этом не показывают общее признаковое описание объектов, которое повлекло это сходство. В этом смысле методы на основе бикластеризации — идеальные кандидаты для поиска сообществ по интересам (общим идеям, проблемам и т. п.), а для анализа триадических данных традиционные методы кластер-анализа неприемлемы. Например, для анализа фолксономических данных (данных вида пользователи-теги-

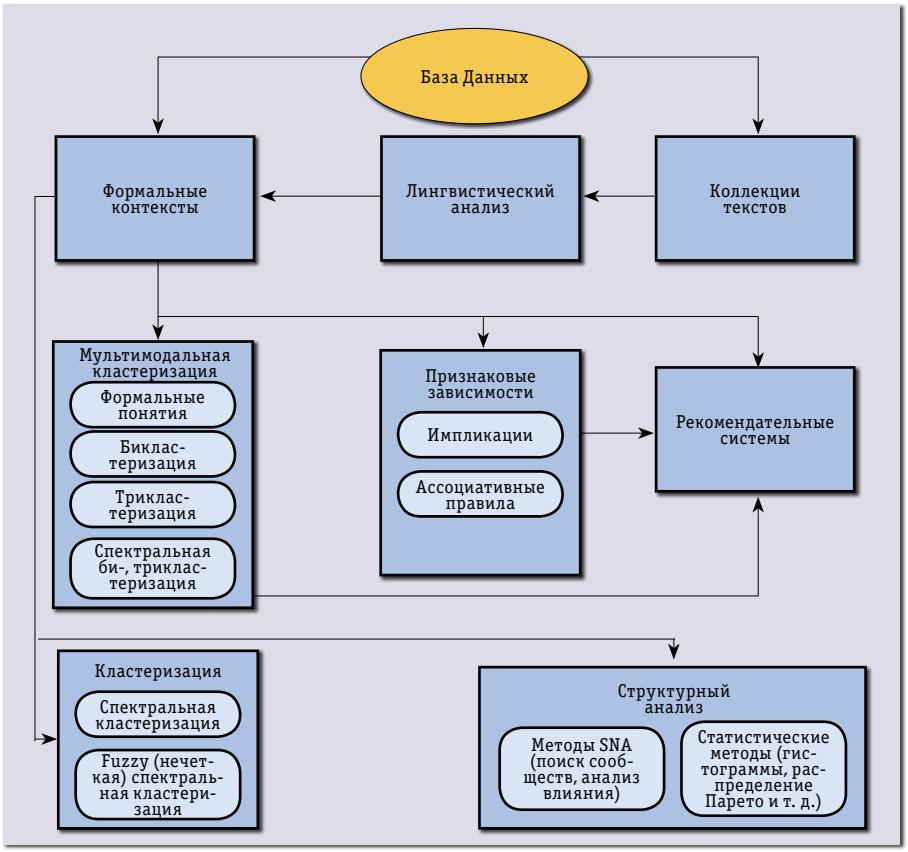


Рис. 1. Схема анализа данных в системе CrowDM

ресурсы) в системах совместного пользования ресурсами успешно применяются методы трикластеризации, в том числе и триадический анализ формальных понятий. Например, у трикластера существует простая геометрическая интерпретация — это параллелепипед в данных, в котором почти все ячейки заполнены. Например, в формальном триконтексте «Пользователи, Теги, Ресурсы» обнаруживается только достаточно плотный трикластер ({Антон, Алена, Настя}, {гаджет, устройство, смарт-книжка}, {Что появится в офисе СБ-21? Что изменится в офисе СБ-21? Уникальная услуга Сбербанк-21}) с 89% заполненных ячеек. Такие плотные трикластеры при разумном выборе порога плотности являются средством описания трехкомпонентных тематических сообществ пользователей. Очевидно, что в этом примере было выявлено тематическое сообщество пользователей, обсуждающих с помощью одинаковых ключевых слов использование «умных» устройств при решении трех различных задач проекта. Очевидно также, что такое сообщество — потенциальный кандидат для образования команды.

Признаковые зависимости (импликации и ассоциативные правила) хотя и не позволяют сделать полноценные выводы о причинно-следственных связях, но дают возможность понять, что в случаях, когда происходят события A, также происходили события B, и это удобно записывается в виде правила A->B. Анализ сходства пользователей по поведению и профилю активности делает возможным применение рекомендательных алгоритмов внутри платформы. А статистические методы, такие как анализ распределений и средних значений, помогают лучше исследовать характер активности пользователей и построить их типологию.

Основное назначение системы CrowDM (рис. 1) состоит в предварительной обработке данных коллаборативной платформы Witology, приведении их к объектно-признаковому виду (формальный контекст) и последующем анализе различными методами. Также в систему встроен блок рекомендательных систем, предназначенный для их тестирования.

АНАЛИЗ ФОРМАЛЬНЫХ ПОНЯТИЙ И КЛАСТЕРИЗАЦИЯ

Главными действующими лицами в краудсорсинговых проектах являются пользователи (участники проекта), которых можно рассматривать как объекты анализа. Каждый объект может обладать (или не обладать) определенным набором признаков: темы, в обсуждении которых пользователь принимал участие; идеи, которые он выдвигал или за которые голосовал; другие пользователи, с которыми он взаимодействовал.

Поиском менее «строгих» групп занимается бикластеризация. Например, объект может не только обладать или не обладать каким-то признаком, а обладать им в той или иной степени. Другой вариант определения бикластера — это формальное понятие с «дырками», то есть с некоторым количеством отсутствующих пар объект-признак.

После формирования формальных контекстов помимо построения формальных понятий, бикластеров и их иерархий можно искать и признаковые зависимости (импликации и ассоциативные правила), то есть утверждения вида «если объект обладает одним множеством признаков, он также обладает и другим». Далее, с помощью таких правил или других известных методов (например, коллаборативной фильтрации на основе сходства по пользователям или признакам, предпочтениям пользователя и истории его поведения на проекте) можно сформировать подсказки, что еще ему могло бы быть интересным в проекте (идеи, люди).

Для проведения первых двух экспериментов с данными коллаборативной платформы Witology в системе CrowDM были отобраны формальные контексты, в которых в качестве объектов выступают пользователи, а в качестве признаков —

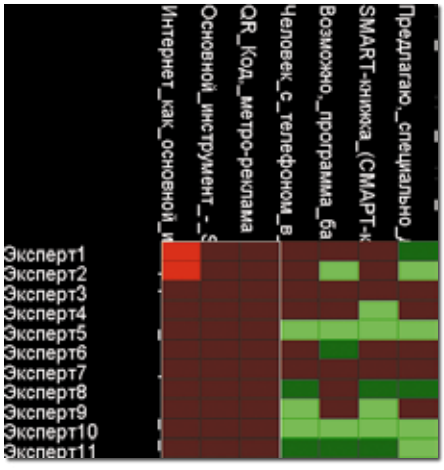


Рис. 2. Пример бикластера из 11 экспертов, обсуждающих общие темы. Цвет ячейки характеризует интенсивность вклада конкретного пользователя в данную проблему

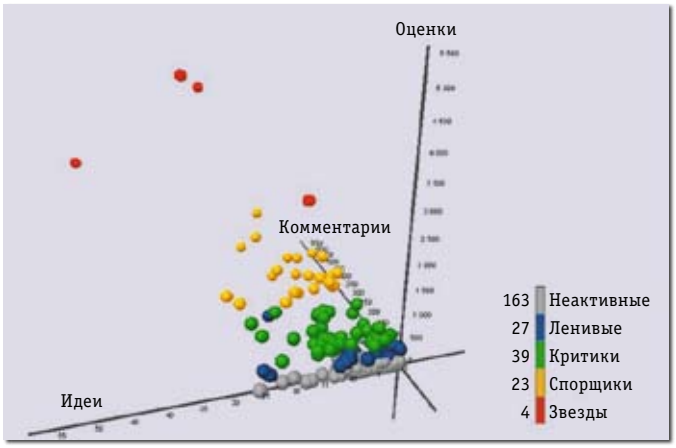


Рис. 3. Результат кластеризации исходных данных

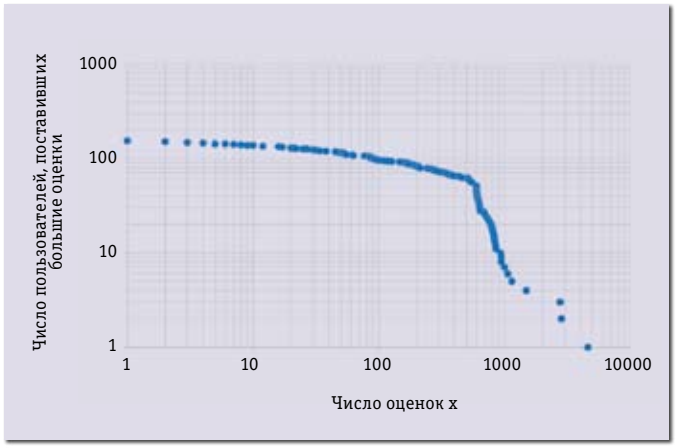


Рис. 4. Распределение пользователей по количеству оценок

идеи, предлагаемые в рамках одной из пяти тем проекта («Сбербанк и частный клиент»). Из всех идей были также отобраны те, которые сумели продержаться до одного из последних этапов проекта. При формировании контекста полагалось, что объект «пользователь» обладает признаком «идея», если данный пользователь внес любой вклад в обсуждение идеи: является ее автором, комментировал идею, выставил ей оценку или комментариям к ней. Таким образом, найденные формальные понятия вида (U,I), где U — множество пользователей, I — множество идей, соответствуют так называемым эпистемическим сообществам (сообществам экспертов или сообществам по интересам) из множества людей U, которые интересуются множествами идей I. Таким образом, было найдено более 100 групп пользователей, а также построена их таксономия в виде решетки понятий.

Эксперименты проводились с помощью программ ConExp для анализа данных на основе решеток формальных понятий (conexp.sourceforge.net) и Meud. После этого было проведено несколько экспериментов по бикластеризации в системе анализа данных генной экспрессии BicAT (Biclustering Analysis Toolbox). В объектно-признаковой таблице, отображаемой в интерфейсе BicAT, строки соответствовали пользователям, столбцы — идеям в рамках указанной темы, в обсуждении которых пользователи принимали участие. Цвет ячейки на пересечении соответствующей строки и столбца отражает интенсивность вклада конкретного пользователя в данную проблему (рис. 2). Данный инструмент позволяет выявлять все возможные группы пользователей по тематике обсуждений, что предоставляет удобную навигацию по найденным сообществам (бикластерам) и тем самым позволяет избежать перестановки строк и столбцов исходной таблицы в уме.

Под вкладом пользователя понимается взвешенная сумма числа его комментариев к этой идее и количества оценок, при этом учитывается, является ли данный человек автором идеи. После дискретизации данных (0 — нулевой вклад, 1 — ненулевой) к ним был применен алгоритм бикластеризации BiMax из пакета BicAT, который нашел несколько крупных бикластеров. Поскольку одной из задач проведения краудсорсинговых проектов является поиск людей со схожими идеями, то наиболее интересен был самый крупный (по количеству пользователей) бикластер из 11 пользователей, в то время как остальные содержали в среднем по четыре-пять пользователей (с ограничением на количество идей в бикластере строго больше двух).

ТИПОЛОГИЯ ПОЛЬЗОВАТЕЛЕЙ

Следующей задачей было выявление типов пользователей согласно их активности. Такая типология важна не только для того чтобы понять, какова доля активных участников краудсорсинговых проектов, но и чтобы найти некие закономерности для разных типов пользователей, которые позволят фасилитаторам проекта адекватно стимулировать их активность.

Изначально имелось три выборки (генерация идей, комментирование, оценивание), каждая из которых состояла из 504 участников проекта «Сбербанк-21». Далеко не все они что-либо создавали или оценивали, поэтому размеры выборок без нулевых значений оказались гораздо меньше количества всех участников проекта. Интерпретация результатов кластеризации по трем параметрам (количество созданных идей, комментариев и выставленных оценок) позволила получить результаты, представленные на рис. 3.

Чтобы более полно увидеть картину активности оценивания в проекте, было построено несколько графиков распределения оценок — например, график из рис. 4 отображает кумулятивное число пользователей, выставивших больше определенного количества оценок за весь проект.

***

Результаты проведенных экспериментов позволяют утверждать, что разрабатываемая методология полезна для анализа данных краудсорсинговых платформ и коллаборативных систем пользования ресурсами. Система CrowDM позволила применить разработанные модели для выявления скрытых тематических сообществ, а также провести анализ активности пользователей платформы Witology и построить их типологию. Предложенные в проекте модели и методы могут быть полезны при анализе данных социальных сетей, систем совместного пользования ресурсами и рекомендательных коллаборативных сервисов. ■

Дмитрий Игнатов (dignatov@hse.ru) — доцент НИУ ВШЭ,
Александра Каминская (alexandra.kaminskaya@witology.com) — аналитик компании Witology,
Андрей Константинов (andrey.v.konst@gmail.com) — научный сотрудник лаборатории интеллектуальных систем и структурного анализа НИУ ВШЭ (Москва).