# ИСПОЛЬЗОВАНИЕ ТОЧЕК ПЕРЕХОДА В ПРЕДУГАДЫВАНИИ СВЯЗЕЙ В СОЦИАЛЬНЫХ СЕТЯХ

**Максим КУКЛЕВ, гр. TI-164**

Технический университет Молдовы

***Аннотация:*** *в данной работе проводится исследование использования инструмента статистики, известного как точки перехода, в решении задачи предсказания связей в социальных сетях рассматриваются упомянутый и классические методы решения. Помимо этого, приведены практические результаты из определённых источников.*

***Ключевые слова:*** *социальная сеть, ссылка, точка перехода, теория графов, кластеризация, статистика.*

Социальные сети стали неотъемлемой частью общества XXI века. Многие бизнес-проекты, компании, государственные учреждения тесно связаны с данным средством массовой информации, которое в отличие от других за последние десятилетия стало интерактивным. Казалось бы, использование данного изобретения, которое на сегодняшний день является ярчайшим представителем среды всемирной компьютерной связи Интернет, демонстрирует одни лишь преимущества, но это совсем не так. Публичность, общедоступность, гениальная простота, выражающаяся в связи миллиардов людей в режиме онлайн, влекут за собой и последствия, пагубно влияющие как на персональную конфиденциальность секретных информационных материалов, так и временами государственную.

Данная проблема порождает задачу, известную как *задачу предугадывания ссылок в соцсетях* (в оригинале – *Link Prediction Problem),* концентрирующуюся на предотвращении нежелательных событий с помощью отслеживания рассматриваемых страниц в Интернете. К данной задачи остаются применимыми следующие методы решения, где за анализируемый объект берётся сама социальная сеть, представленная в виде абстрактного графа:

- методы соседства вершин графа;
- методы группирования путей графа;
- методы высшего порядка.

Каждый из перечисленных методов принимает следующие аксиомы:

- Главные значения в любом алгоритме – вершины графа и вес рёбер между ними.
- Существует огромное множество невыявленных рёбер, которые и являются искомыми связями.

Подход классических методов можно рассмотреть на примере *кластеризации*, или *кластерного анализа* – одного из методов высшего порядка. Предположим, имеется граф с очень большим количеством вершин. Вышеупомянутый алгоритм будет анализировать граф и отбрасывать вершины, которые расположены относительно сингулярно – далеко от остальных вершин. Соответственно, с каждой итерацией граф будет постепенно преобразовываться в скопления вершин, как на приведённом изображении (рис. 1 [1]).

Алгоритм сходится в случае любой метрики, но зависим от исходного положения центральных точек кластеров в пространстве. В таких случаях помогают последовательные запуски с последующим усреднением результативных кластеров. В этих уменьшенных группах и будут наблюдаться самые посещаемые веб-страницы, выкладывающиеся в определённую схему популярной, актуальной информации, которой обмениваются на пространстве социальных сетей.

В начале 2000-ых было предложено использовать в данной задаче *точки перехода*– термин, пришедший из статистики, обозначающий переход от одного значения к другому в распределении вероятностей. В контексте задачи термин будет обозначать момент возникновения новых связей в течение времени. Таким образом, основываясь на частоте и областях появления новых связей, , становится возможным выделять новые закономерности.

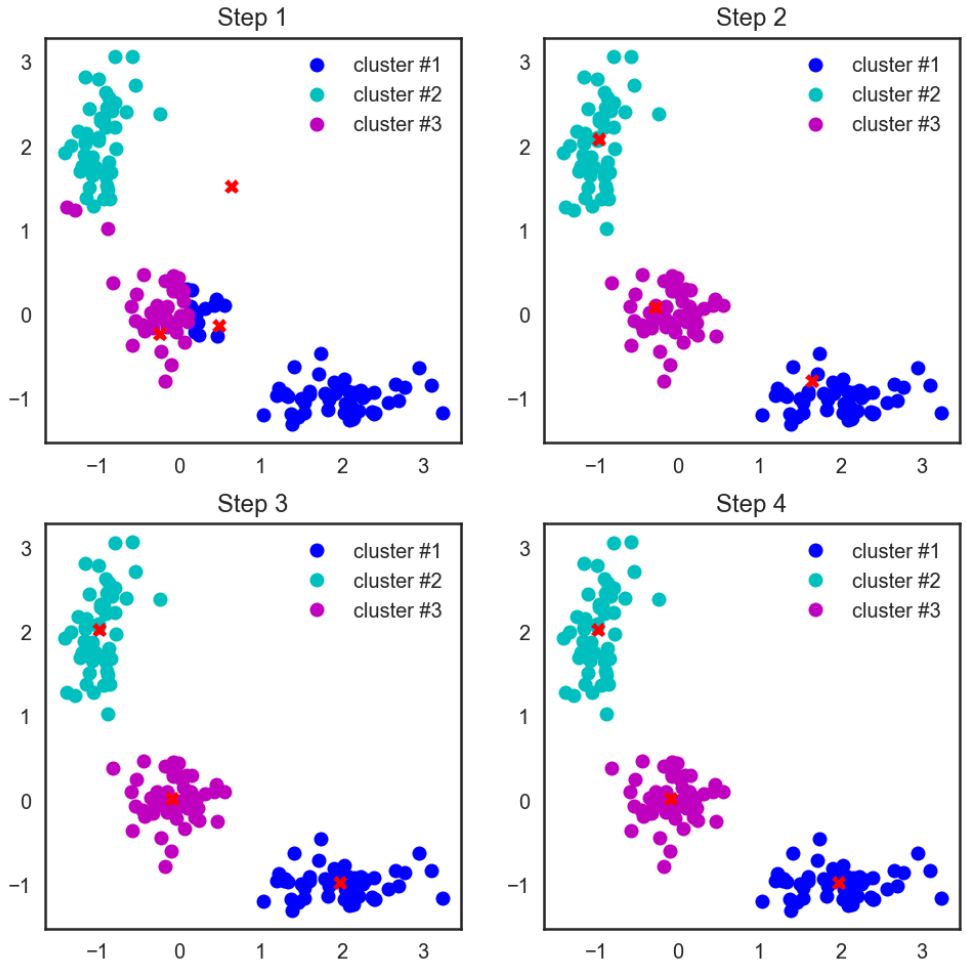


Рис. 1. Графическое представление алгоритма кластеризации

Метод точек перехода использует следующие метрики:

- *плотность графа* – мера, описывающая насколько интенсивно граф заполнен рёбрами;
- *средняя промежуточная центральность* (англ. *average betweenness centrality*)– показывает насколько часто рёбра проходят через рассматриваемую вершину;
- *средняя близостная центральность* (англ. *average closeness centrality)* – демонстрирует насколько быстро можно добраться из выбранной вершины во все остальные.

Если говорить об областях применения метода, примечательны такие области, как бизнес-аналитика и маркетинг, так как посещаемость страниц активно рассматривается на таких пространствах, как интернет-магазины, а также рекламные блоки в социальных сетях; большую роль данный метод играет в продумывании и построении архитектуры и функционала социальных сетей для повышения популярности её использования среди пользователей. Но важнейшей областью, несомненно, остаётся криминалистика, когда необходимо отследить и выявить скрытую связь сообщения между членами преступных группировок. В качестве примера можно привести исследование Йена МакКаллоу, Мэттью Уэбба, Джона Грэма, Кэтлин Карли, Дэниэла Хорна, в сотрудничестве с Военной Академией США, Исследовательским институтом армии США и Университетом Карнеги Меллон *«Точки перехода в социальных сетях»* (ориг. *«Change Detection in Social Networks»*). В нём было проведено исследование связи террористической сети Аль-Каида в среде интернет в динамическом прогрессировании в течение нескольких лет. По полученным данным удалось построить график зависимости активности в подпольной сети от времени (рис. 2 [2]).

Становится заметным, что пик пришёлся на начало XXI века, когда международная обстановка между «Аль-Каидой» и США накалилась практически до предела. Затем виден резкий спад, однако, чем же он обусловлен, на первый взгляд, непонятно. Обратившись к историческим справкам, ситуацию

можно легко прояснить – 11 сентября 2001 года «Аль-Каида» был совершила серию самоубийственных террористических актов в Нью-Йорке. Впоследствии секретные службы Соединённых Штатов занялись поисками и внедрением в коммуникацию террористов, перехваткой сообщений, а также заглушкой их контакта.

Автором Павлом Сулимовым с портала ХабраХабр, основателем компании “Crocode”, более известного под никнеймом @crocodeinc, был выполнен эксперимент с данным методом над хостингом фотографий и видео Flickr. За основные измерения он взял коэффициент Жаккара, сумму локальных коэффициентов кластеризации (метрика, основанная на свойствах вершин), значения трех основных мер центральности узла (для каждой пары вершин сумма degree centrality, сумма closeness centrality и сумма betweenness centrality) и значение кратчайшего расстояния между вершинами, взятое со знаком минус.

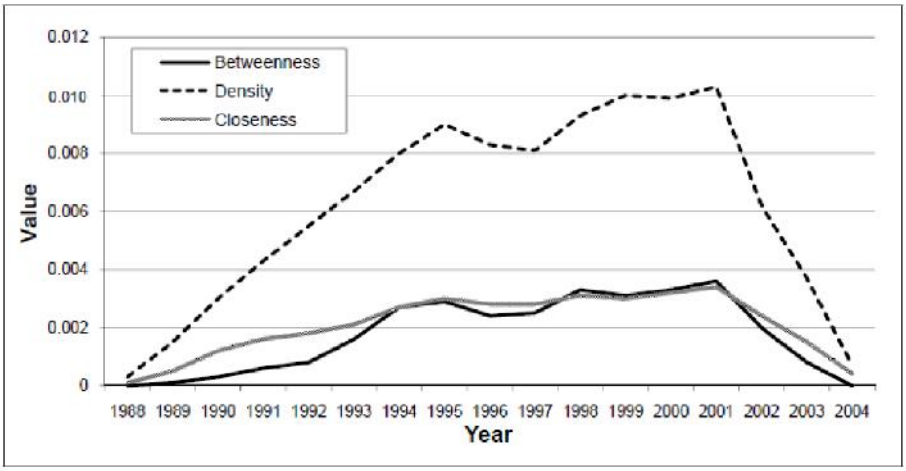


Рис. 2. График зависимости активности группировки Аль-Каида в сети на протяжении десятилетий

Автор предположил, что вероятность соединения пары вершин может зависеть не только от параметров и состояний в предыдущий момент времени, но и в момент некоторых временных задержек и лагов, так как, если наблюдается рост показателя какого-то классификатора со временем, значит, соединение 2 вершин на каждом шаге становится всё более вероятным. Далее отмечается включение характеристик, ответственных за обнаружение точек перехода в сети, но предпочтение отдаётся абсолютным значениям плотности, средней промежуточной и близостной центральностей. Вводится булевая переменная link, отражающая наличие, или же отсутствие связи между парой рассматриваемых вершин (1 и 0 соответственно). Решается задача двухклассовой классификации посредством метода случайного дерева, одним из алгоритмов машинного обучения кластеризации, и в результате были получены значения, продемонстрированные на матрице ошибок (рис. 3 [3]).

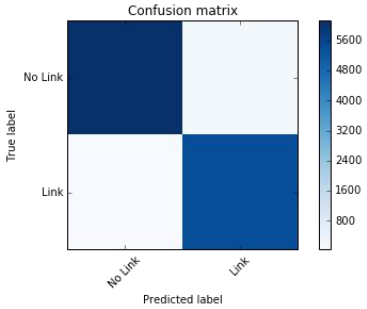


Рис. 3. Матрица ошибок со значениями link и полученными экспериментальными данными

Проанализировав полученные данные, можно постановить, что использованный в опыте метод соответствует ожиданиям – согласно приведённой шкале, подавляющее большинство выходных значений совпали с предполагаемыми. Ожидания насчёт значений, в которых связей не предполагалось (в матрице пересечение no link и predicted label), подтвердились (они попали в диапазон пересечения

no link и true label). Аналогично получилось и с ожидаемыми связями. В результате эксперимента Павел Сулимов привёл следующие выводы:

- самые ценные показатели – average close и close (меры того, насколько быстро можно дойти из одной вершины в остальные) способны предсказывать именно будущие, а не настоящие связи;
- эти показатели были рассчитаны не на основе собственных свойств вершин, а на основе показателей всего графа;
- значимость средних близостных и центральных центральностей говорят о том, что гипотеза наличия точек перехода подтверждена на практике;
- введение точек перехода способствует более точному предсказанию классов связей.

Рассмотрев тему с разных сторон, в заключение с уверенностью можно сказать, что алгоритм представляет собой успешную модель работы с прогнозированием. Он является достаточно эффективным и актуальным во многих сферах современного мира и сочетает в себе элементы статистики, теории вероятности и машинного обучения. За последнее десятилетие метод точек перехода стал ключевым в разработке веб-технологий и мониторинге социальной кибербезопасности.

**Библиография**


1. @libfun, Открытый курс машинного обучения. Тема 7. Обучение без учителя: PCA и кластеризация [Электронный ресурс]. – Режим доступа: https://habrahabr.ru/company/ods/blog/325654/
2. Ian McCulloh, Matthew Webb, John Graham, US Military Academy, Kathleen Carley, Carnegie Mellon University, Daniel B. Horn, US Army Research Institute, Change Detection in Social Networks [Электронный ресурс]. – Режим доступа: http://www.casos.cs.cmu.edu/publications/papers/TR%201235.pdf
3. Сулимов П., @crocodeinc, Предсказание связей в социальных сетях: используем точки перехода [Электронный ресурс]. – Режим доступа: http://habrahabr.ru/post/341718
4. David Liben-Nowell, MIT, Jon Kleinberg, Cornell University, The Link Prediction Problem for Social Networks [Электронный ресурс]. – Режим доступа: https://www.cs.cornell.edu/home/kleinber/link-pred.pdf
5. Dr. Wayne A. Taylor, Change-Point Analysis: A Powerful New Tool for Detecting Changes [Электронный ресурс]. – Режим доступа: http://www.variation.com/cpa/tech/changepoint.html
6. Хажоян А.В., Предсказание ссылок в социальных сетях [Электронный ресурс]. – Режим доступа: https://dspace.spbu.ru/bitstream/11701/4118/1/diploma.pdf