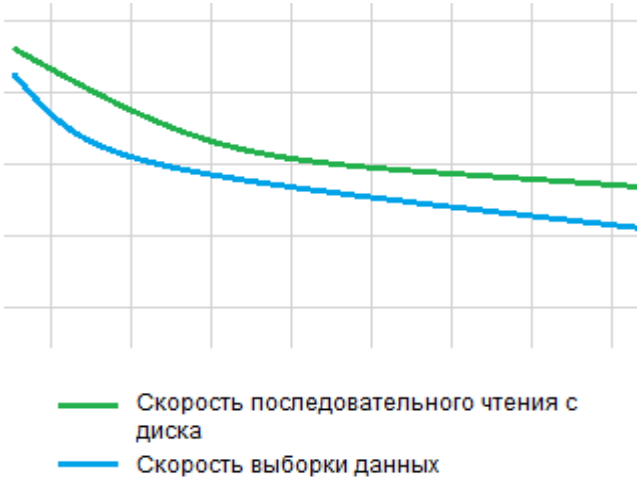

Рис. 4 – Сравнение скоростей линейного чтения с диска и выборки данных после оптимизации

Это решение позволяет достигать скорости выборки данных сопоставимой со скоростью линейного чтения, и не зависит от используемых СУБД.

Список использованных источников:

1. Моррисон, Алан. Большие Данные: как извлечь из них информацию/ АланМоррисон // Ежеквартальный журнал– Москва, 2010. – 31 с.
2. Свободная общедоступная многоязычная универсальная энциклопедия [Электронный ресурс], http://en.wikipedia.org/wiki/Hard_disk_drive
3. Свободная общедоступная многоязычная универсальная энциклопедия [Электронный ресурс]: https://ru.wikipedia.org/wiki/IOPS.

МАКСИМИЗАЦИЯ ВЛИЯНИЯ В СОЦИАЛЬНЫХ СЕТЯХ

Белорусский государственный университет информатики и радиоэлектроники
г. Минск, Республика Беларусь

Лычковский А. В.

Волорова Н. А. – к-т. техн. наук, доцент

Социальная сеть отношений между людьми играет важную роль в распространении информации, идей и влияния среди ее членов. Новое мнение или инновация, которые появляются в мире, могут стать невостребованными или внести значительный вклад в развитие общества. Для того, чтобы понять степень принятия новых идей, нужно понять какова динамика их распространения в социальной сети. Эта динамика, в свою очередь, определяется степенью влияния одного объекта на другой, так что решение одного объекта приводит к принятию другим схожего решения.

Процессы, происходящие в социальных сетях, имеют долгую историю исследования. Некоторые из ранних исследований касались использования новых медицинских препаратов и инноваций в сфере сельского хозяйства [1, 2]. Также исследовались проблемы распространения информации о новом продукте при использовании «вирусной рекламы» [3,4,5]. Схожие процессы могут быть найдены и в инженерных сетях, например, проблема каскадного отключения энергетических систем [6].

Последние работы в этой области, касаются проблемы маркетинга [5]. Предположим, что мы имеем данные о некоторой сети, включая оценку степени влияния одного объекта на другой. В этой сети мы хотим рекламировать новый продукт таким образом, чтобы максимальная ее часть приняла продукт. Для этого можно воспользоваться «вирусной рекламой». Ее суть в том, чтобы вовлечь некоторое начальное множество «влиятельных» объектов из социальной сети, продемонстрировать им продукт, например, раздавая бесплатные образцы. Далее ожидается, что эти объекты будут делиться своими мнением с своими друзьями, потом друзья со своими друзьями и так далее. При такой формулировке возникает вопрос, как выбрать начальное множество «влиятельных» объектов таким образом, чтобы как можно большая часть социальной сети приобрела продукт.

Математическая модель описанной выше задачи может быть представлена в следующем виде. Имеется ориентированный взвешенный граф $G = (V, E)$ и модель M процесса распространения идеи в G. Для любого подмножества $A \subseteq V$ пускай $\delta(A)$ определяет ожидаемое число объектов, которые приняли идею, если A было начальным множеством. Таким образом входными значениями для задачи будут являться граф $G = (V, E)$ и натуральное число K, которое определяет максимальное число элементов в

начальном множестве. Выходным значением будет подмножество A,, такое что $A \subseteq V, |A| \leq K$ для которого $\delta(A)$ максимально.

В описании математической модели упомянута модель распространения идеи M. Рассмотрим две базовые и наиболее часто используемые из них: *линейного порога* и *независимого каскада.* Обе модели рассматривают вершины графа G как активные и неактивные. *Активными* считаются вершины представляющие объекты, которые приняли идею, *неактивными*, которые не приняли. Для упрощения понимания процесса рассмотрим только положительный вариант развития, который подразумевает то, что объект однажды принявший идею никогда от нее не отказывается. Таким образом, в некоторый момент времени достаточное количество друзей объекта v может принять идею, что заставит сам объект v изменить свое мнение и влиять на друзей не принявших идею. Суть модели *линейного порога* в том, что каждый объект имеет некоторый *порог* [7], после которого он принимает идею. В этой модели на объект v оказывает влияние каждый соседний объект w, согласно весу $b_{v,w}$ такому, что $\sum_{w \text{ сосед } v} b_{v,w} \leq 1$. Динамика такого процесса развивается следующим образом. Для каждого объекта v существует порог θ_v из интервала $[0,1]$. Для данного набора активных объектов A_0(все остальные объекты неактивны), распространение может быть описано по шагам: на шаге t, все объекты, которые были активны на шаге $t-1$, остаются активными, а также становятся активными объекты для которых общее влияние их соседей больше θ_v: $\sum_{w \text{ активный сосед } v} b_{v,w} \geq \theta_v$. Другая модель описывающая тот же процесс оперирует терминами теории вероятности и называется моделью *независимого каскада*. Также как и предыдущая модель она имеет начальный набор активных объектов A_0 и развертывается дискретно во времени. Когда объект v становится активным на шаге t, он получает единственный шанс активировать каждого соседа w. Объект реализует свой шанс с вероятностью $p_{v,w}$, которая задается изначально и не меняется в течение времени. Если объект w имеет несколько ново активированных соседей, каждый их них получает шанс последовательно активировать w. Если v реализует шанс, то w становится активным на $t+1$ шаге. Но независимого от результата, v не выполняет попыток активации w на последующих шагах. Процесс активации в обеих моделях останавливается, если на шаге t не было активировано новых объектов.

Ранние подходы к решению этой задачи были основаны на вероятностных алгоритмах [5]. Затем было предложено рассматривать задачу как дискретную оптимизационную [8]. Было доказано, что оптимизационная задача является NP сложной. Также показано, что применение жадного алгоритма к рассмотренным моделям гарантирует, что найденное $\delta(A)$ отличается от оптимального в $(1-\frac{1}{e})$ раз. Суть алгоритма в том, что он итеративно формирует результирующее множество S, такое, что $\delta(S)$ максимально, и $|S| = K$. В начале работы алгоритма $S = \emptyset$, далее, на каждой итерации, выбирается объект v, $v \in A \backslash S$ такой что $\delta(S \cup \{v\})$ максимально, этот объект добавляется в S. Алгоритм останавливается, когда $|S| = K$. Однако жадный алгоритм имеет серьезный недостаток – производительность. Вычисление $\delta(\{v\})$ на каждом шаге требует больших затрат. Эту проблему пытаются решить с помощью оптимизации базового алгоритма [9, 10].

Псевдокод жадного алгоритма представлен ниже
1: $S = \emptyset$
2: For i = 1 to K do
3: Выбрать v: $\delta(S \cup \{v\}) - \delta(S)$ максимально
4: $S = S \cup \{v\}$
5: return S

Существуют также и другие алгоритмы, которые используют специфические модели M.

Таким образом рассмотрены наиболее известные работы в области максимизации влияния, историческое развитие проблемы, а также ее прикладное значение. Из описанного выше видно, что интерес для исследования представляют алгоритмы максимизации $\delta(A)$. Для этого следует проделать работу в двух направлениях: оптимизация существующих алгоритмов, анализ моделей M и представление их в виде, удобном для написания эффективного алгоритма.

Список использованных источников:

1. J. Coleman, H. Menzel, E. Katz. Medical Innovations: A Diffusion Study Bobbs Merrill, 1966.
2. T. Valente. Network Models of the Diffusion of Innovations. Hampton Press, 1995
3. F. Bass. A new product growth model for consumer durables.Management Science 15(1969), 215-227.
4. J. Brown, P. Reinegen. Social ties and word-of-mouth referral behavior. Journal of Consumer Research 14:3(1987), 350-362
5. P. Domingos, M. Richardson. Mining the Network Value of Customers. Seventh International Conference on Knowledge Discovery and Data Mining, 2001
6. C. Asavathiratham. The Influence Model: A Tractable Representation for the Dynamics of Networked Markov Chains. Ph.D. Thesis, MIT 2000
7. M. Granovetter. Threshold models of collective behavior. American Journal of Sociology 83(6):1420-1443, 1978.
8. D. Kempe, J. M. Kleinberg, and É. Tardos. Maximizing the spread of influence through a social network. In Proceedings of the 9th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, pages 137–146, 2003.
9. J. Leskovec, A. Krause, C. Guestrin, C. Faloutsos, J. VanBriesen, and N. S. Glance. Cost-effective outbreak detection in networks. In Proceedings of the 13th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, pages 420–429, 2007.
10. A. Goyal, W. Lu, and L. V. S. Lakshmanan. Celf++: optimizing the greedy algorithm for influence maximization in social networks. In WWW, pages 47–48, 2011.