ПОСТРОЕНИЕ АССОЦИАТИВНОГО РЯДА ХЭШТЕГОВ С ИСПОЛЬЗОВАНИЕМ СЕТИ СОВМЕСТНОЙ ВСТРЕЧАЕМОСТИ И ВЕКТОРНОГО ПРЕДСТАВЛЕНИЯ ХЭШТЕГОВ

Макрушин С.В., Блохин Н.В.
Финансовый университет при Правительстве Российской Федерации, г. Москва
SVMakrushin@fa.ru

Решение задач навигации в сетях, таких, как поиск кратчайшего пути, является одной из классических задач, начиная с зарождения теории графов и в различных постановках актуально практически для всех предметных областей. В предложенной работе рассматривается задача построения кратчайшего ассоциативного ряда между двумя заданными хэштегами в сети совместной встречаемости хэштегов, построенной для сообщений из социальной сети Инстаграм. Построение такого пути может быть практически интересно, например, для рекомендации пользователям хэштегов, связанных с двумя изначально заданными хэштегами. Кратчайший ассоциативный ряд может рассматриваться как кратчайший путь между двумя узлами сети совместной встречаемости хэштегов, при этом веса связей или дополнительные критерии, или ограничения при построении пути в том или ином виде должны формализовать условие о наличии сильной смысловой связи между соседними узлами построенной цепочки хэштегов.

Для построения сети совместной встречаемости хэштегов Инстаграм был собран корпус из 14,6 миллионов сообщений, содержащих хэштеги. Хэштеги являются узлами созданной сети и для каждого из них хранится количество его упоминаний в собранном корпусе сообщений. Если в рассматриваемом корпусе два хэштега совместно встречались более чем в двух сообщениях, то между ними в сети установлена связь, для которой хранится количество совместных упоминаний этих хэштегов. В рамках работы была построена сеть из 6,13 миллионов узлов (хэштегов) и 63,89 миллионов связей между ними. Распределение степеней узлов для сети приведено на рис. 1.а. Сбор и предварительная обработка данных были реализованы на языке программирования Python, хранение сети в графовой базе данных ArangoDB.

Для больших сетевых структур из реального мира, таких как построенная сеть хэштегов, вычислительная сложность точного решения задачи построения кратчайшего пути и, как следствие, построения ассоциативного ряда может оказаться неприемлемой. В рамках исследования был предложен подход к преобразованию сети совместной встречаемости, который позволил снизить вычислительную сложность построения ассоциативного ряда с помощью построения глобально оптимального кратчайшего пути, а также основанный на преобразованной сети вычислительно эффективный алгоритм построения ассоциативного рада, использующий только локальную информацию о сети и информацию о векторном представлении хэштегов, полученном с помощью алгоритма word2vec [1].

Как видно из рис. 1 хвост распределения степеней узлов в рассматриваемой сети близок к степенному закону распределения. Таким образом, в сети имеется значительное количество «хабов», обладающих десятками и сотнями тысяч связей. Анализ хэштегов-хабов показал, что 80% всех случаев совместных упоминаний хабов обычно обеспечиваются всего несколькими десятками хэштегов, сильно связанных с данным хабом. Именно такие ключевые связи в сети демонстрируют наибольшую смысловую взаимосвязь между хэштегами. Избавившись от большого количества неважных связей в сети, можно как снизить вычислительную сложность реализации построения ассоциативной цепочки, так и гарантировать наличие сильной смысловой связи между соседними узлами.

Рассмотрим возможность перехода от хэштега i к хэштегу j в ассоциативном ряде как событие с вероятностью $p_{ij} = \frac{C_{ij}}{T_i}$, где C_{ij} – количество совместных упоминаний хэштегов i и j, а $T_i = \sum_{j \in N(i)} C_{ij}$ – суммарное количество совместных упоминаний хэштега i со всеми соседними хэштегами. В рассматриваемой сети хэштегов мы отбросили наименее важные связи, которые в сумме дают не более 20% вероятности перехода для каждого из хэштегов. Нужно отметить, что так как $p_{ij} \neq p_{ji}$, то в результате модификации была получена сеть с ориентированными связями, которых в нашем случае оказалось 61,88 миллиона. График распределения исходящих степеней узлов в логарифмическом масштабе представлен на рис. 1.б. На графике видно, что модификация позволила снизить количество исходящих связей для «хабов» почти на порядок, что может принципиально снизить вычислительную сложность поиска кратчайшего пути.

На основе определения p_{ij} – вероятности перехода от одного хэштега к другому – мы сформулировали требование к оптимальному ассоциативному ряду хэштегов $(h_l, h_2, \cdots, h_n)$, связывающему хэштег $i = h_1$ с хэштегом $j = h_n$, как: $\max_{(h_1, h_2, \cdots, h_n) \mid h_1 = i,\, h_n = j} \prod_{l=1}^{n} p_{h_l h_{l+1}}$. Тогда оптимальный ассоциативный ряду будет достигаться на кратчайшем пути в сети с весами ориентированных связей $-ln\, p_{ij}$. При такой формулировке условие о смысловой связи соседних хэштегов ассоциативного ряда достигается как за счет весов связей, так и за счет отсутствия в сети связей с низким рангом вероятности p_{ij}. Примеры ассоциативных рядов, полученные при помощи построения глобального кратчайшего пути для модифицированной сети хэштегов приведены в Таблице 1.

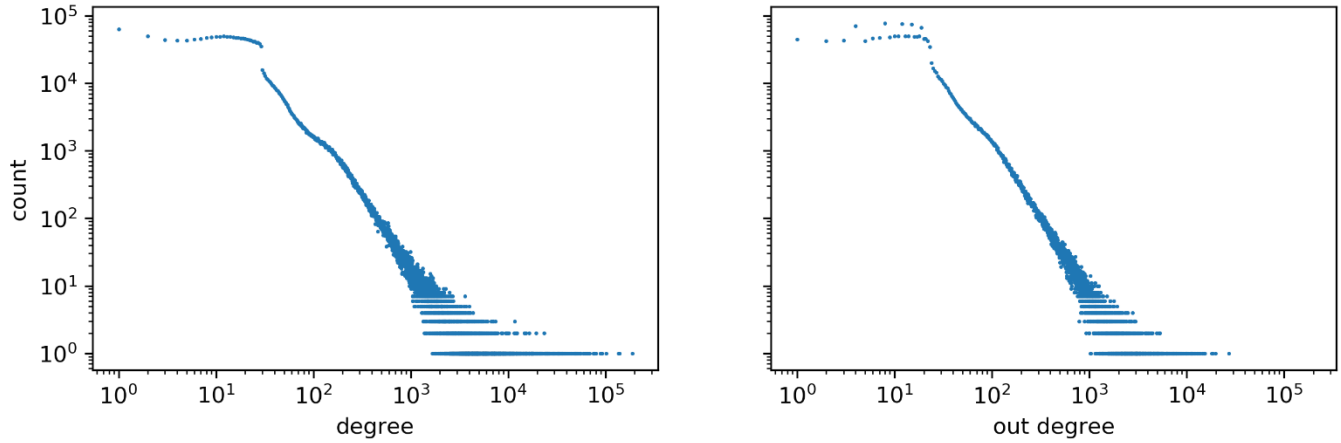
Если для узлов сети доступна метрика, связанная с расстоянием между узлами в сети, то с ее помощью можно построить квазиоптимальный кратчайший путь между узлами сети, например, с использованием жадного алгоритма «близорукого поиска», опирающегося только на локальную информацию о сети (см., например [2]). Преимуществом такого подхода является низкая вычислительная сложность поиска пути.

При необходимости метрику для узлов сети можно построить, опираясь только на топологию самой сети [3, 4]. Например, алгоритм node2vec [4] при помощи случайного блуждания по сети порождает последовательность упоминаний узлов сети и применяет к ней алгоритм word2vec для получения векторного представления узлов сети. Однако для случая сетей совместной встречаемости векторное представление узлов легко получить напрямую, применив алгоритм word2vec к исходному множеству сообщений, по которым была построена сеть. С помощью библиотеки gensim для собранного корпуса сообщений из Инстаграм нами было построено векторное представление для каждого хэштега. Благодаря этому каждый хэштег характеризуется 300-мерным вектором, что позволяет, используя косинусную меру, очень просто определять расстояние между любыми двум хэштегами в сети.

Для построения ассоциативной цепочки на основе локальной информации о сети был взят жадный алгоритм «близорукого поиска» кратчайшего пути, описанный в [5]. Этот алгоритм был применен на модифицированной сети совместной встречаемости хэштегов с метрикой, порожденной полученным векторным представлением хэштегов. Таким образом, нами предложен оригинальный подход для поиска вычислительно эффективного решения задачи построения ассоциативного ряда на основе совместного использования техники получения векторного представления слов и сети совместной встречаемости терминов. Примеры ассоциативных рядов, полученные этим способом для рассматриваемой сети хэштегов, приведены в Таблице 1.

Литература

1. *T. Mikolov, et al*. Efficient Estimation of Word Representations in Vector Space // arXiv:1301.3781 / 2013
2. *J. Kleinberg*, The Small-World Phenomenon: An Algorithmic Perspective // Proc. 32nd ACM Symp. Theory of Computing / 2000, pp. 163-170
3. *P. Goyal, E. Ferrara*, Graph Embedding Techniques, Applications, and Performance: A Survey. // *Knowl.-Based Syst.* / 2018 #151: pp. 78-94.
4. *A. Grover, J. Leskovec*, node2vec: Scalable feature learning for networks // Proceedings of the 22nd International Conference on Knowledge Discovery and Data Mining, ACM / 2016, pp. 855–864.
5. *J.A. Capitan et al.*, Local-Based Semantic Navigation on a Networked Representation of Information // PLoS ONE 7(8): e43694 / 2012



а. б.

Рис. 1 Распределение степеней (количества связей) для сети совместной встречаемости хэштегов Инстаграм **а.** Для исходной сети **б.** Для степеней исходящих связей модифицированной сети без связей с низкой вероятностью перехода.

Таблица 1. Примеры ассоциативных рядов, построенных с помощью глобального кратчайшего пути и жадного локального алгоритма построения кротчайшего пути с использованием векторного представления хэштегов

Глобальный кратчайший путь	**Локальный алгоритм, использующий векторное представление**
#серьгискристаллами	***#серьгискристаллами***
#москва	#вечерниесерьги
#маркетинг	#ювелирныеукрашения
	#реклама
	#маркетинг
#домвгорах	***#домвгорах***
#летнийотдых	#домвдагомысе
#гражданскиедела	#строимдомвсочи
	#недвижимость
	#гражданскиедела
#мояеда	***#мояеда***
#ставрополь	#правильнокушатьзологздоровья
#пряникнаторт	#пппряники
	#пряники
	#пряникнаторт
#веганекб	***#веганекб***
#скидка	#этносемья
#медицинскийкостюм	#маслоизконопли
	#едаизконопли
	#натурально
	#доктор
	#медицинскийкостюм