Максимов Артур Вадимович
магистр, аспирант
Смолина Светлана Георгиевна
канд. техн. наук, доцент
АНО ВО «Российский новый университет»
г. Москва

СРАВНИТЕЛЬНЫЙ АНАЛИЗ МЕТОДОВ ПРЕДСКАЗАНИЯ СВЯЗЕЙ В ТЕМПОРАЛЬНЫХ СЕМАНТИЧЕСКИХ СЕТЯХ ДЛЯ ВЫЯВЛЕНИЯ ЭВОЛЮЦИИ НАУЧНЫХ КОНЦЕПЦИЙ

Аннотация: *в статье исследуется предсказание связей в темпоральных семантических сетях научных концептов, где узлы – ключевые термины, а связи – их ко-оккурренция в публикациях. Проведено эмпирическое сравнение четырёх методов – Common Neighbors, NMF, Node2Vec и Dyngraph2Vec – на корпусе научных статей за несколько лет. Качество оценивается метриками AUCROC и Average Precision в режиме скользящего контроля по временным срезам. Эксперименты показывают, что темпоральные эмбеддинги статистически значимо превосходят как эвристики, так и статические методы, особенно при выявлении новых связей между ранее далёкими концептами. Результаты дают обоснованную рекомендацию по выбору инструментария для предиктивной наукометрии и формируют базовую линию для дальнейших исследований.*

Ключевые слова: *темпоральные семантические сети, предсказание связей, эволюция научных концепций, ко-оккуренция, бенчмарк, наукометрия, графовые эмбеддинги, анализ цитирования, динамические графы.*

Изучение эволюции научного знания – одна из центральных задач наукометрии. С развитием методов обработки естественного языка и анализа графов все большее распространение получает представление научного дискурса в виде темпоральных семантических сетей, где узлами выступают ключевые концепты, а связи отражают их совместную встречаемость в публикациях. Динамика этих

связей во времени позволяет не только ретроспективно описывать развитие научных направлений, но и формулировать прогнозные гипотезы о зарождающихся трендах.

В основе построения подобных прогнозов лежит задача предсказания связей – оценка вероятности возникновения ребра между ранее не связанными концептами. За последние годы предложен широкий арсенал подходов: от элементарных топологических эвристик до сложных темпоральных моделей на графовых эмбеддингах. Однако практически все эмпирические сопоставления этих методов ограничены социальными, биологическими или коммуникационными сетями. Для семантических сетей научных терминов, отличающихся высокой разреженностью, неоднородной встречаемостью концептов и уникальной динамикой образования новых связей, систематический бенчмарк до сих пор не проводился.

Цель данной работы – провести строгое эмпирическое сравнение четырех репрезентативных методов предсказания связей, относящихся к разным алгоритмическим классам: топологическая эвристика (Common Neighbors), матричная факторизация (NMF), статические графовые эмбеддинги (Node2Vec) и темпоральные эмбеддинги (Dyngraph2vec). Эксперимент выполняется на темпоральной семантической сети, построенной из реального корпуса научных публикаций. Качество методов оценивается по стандартным метрикам AUC-ROC и Average Precision с использованием скользящего контроля по временным срезам.

Новизна работы заключается в первом систематическом бенчмарке указанных классов методов именно на семантических сетях научных концептов, а также в публикации формализованного программного пайплайна в открытом доступе, что обеспечивает воспроизводимость результатов и возможность их расширения.

Задача предсказания связей в сложных сетях привлекает внимание исследователей на протяжении последних двух десятилетий. Наиболее полные на сегодняшний день обзоры [4][7] предлагают таксономию методов, разделяя их на эвристические, основанные на факторизации матриц и основанные на представле-

ниях (эмбеддингах). При этом подавляющее большинство эмпирических сравнений, приводимых в этих обзорах, выполнено на социальных, биологических и коммуникационных графах.

Элементарные топологические эвристики – например, Common Neighbors или Adamic Adar – не нуждаются в обучении и выступают надёжным базовым уровнем (бейзлайном) [6]. Методы матричной факторизации, в частности неотрицательная матричная факторизация, дают возможность выявить латентные факторы, объясняющие структуру связей в статическом срезе сети [5]. Дальнейшим развитием стали графовые эмбеддинги, среди которых Node2Vec [2] остаётся одним из наиболее цитируемых подходов, преобразующих топологию графа в векторное пространство.

Общим ограничением перечисленных методов является пренебрежение временно́й динамикой: все они оперируют единственным статическим снимком сети. Для моделирования эволюции были предложены темпоральные эмбеддинги, в частности Dyngraph2vec [1], который объединяет глубокий автоэнкодер с рекуррентным слоем для предсказания будущих представлений узлов. Отдельные работы [3] показывают, что в ряде случаев статические эмбеддинги способны конкурировать с темпоральными, однако подобные сопоставления не проводились на семантических графах научных терминов.

Таким образом, в литературе отсутствует систематический бенчмарк, который бы в единых условиях сопоставил топологические эвристики, матричную факторизацию, статические и темпоральные эмбеддинги применительно к задаче предсказания связей в темпоральных семантических сетях научных концептов.

Эксперимент проводится на открытом корпусе научных публикаций (ACL Anthology, 2019–2023). Из заголовков и аннотаций статей с помощью алгоритма YAKE извлекаются ключевые термины. Для обеспечения устойчивости удаляются термины, встретившиеся менее чем в трех документах за год, а также общеязыковые стоп-слова. Для каждого года строится неориентированный граф ко-оккурренции: ребро между двумя терминами проводится, если количество предложений, в которых они встретились совместно, не менее 3. В результате

получается последовательность годовых графов G_2019, …, G_2023, образующая темпоральную семантическую сеть. Все графы хранятся в виде матриц смежности. Задача предсказания связей ставится как бинарная классификация. Для временного окна, заканчивающегося годом T, положительными примерами являются пары терминов, впервые образовавшие ребро в графе G_{T+1} и не имевшие связи ни в одном из графов G_1…G_T. Отрицательные примеры – случайно выбранные пары без связи в G_{T+1}, но имеющие хотя бы одного общего соседа в G_T; это исключает тривиально несвязные пары и усложняет задачу. Классы балансируются, чтобы избежать смещения метрик.

Отобраны четыре метода из разных классов:

Common Neighbors (CN) – топологическая эвристика без обучения. Скор p(i, j) равен числу общих соседей вершин i и j в графе G_T.

NMF – неотрицательная матричная факторизация матрицы смежности A_T графа G_T с рангом d = 64. Скор – скалярное произведение векторов-строк матрицы W.

Node2Vec – статические эмбеддинги, обученные на G_T с размерностью векторов d = 128. Скор – косинусное сходство полученных представлений.

Dyngraph2vec – темпоральная модель, объединяющая глубокий автоэнкодер и рекуррентный LSTM-слой. Обучается на всей последовательности G_1…G_T, предсказывая эмбеддинги вершин для момента T+1. Скор – косинусное сходство предсказанных векторов.

Используется скользящий контроль с двумя последовательными окнами: обучение на 2019–2021 и тест на 2022; обучение на 2019–2022 и тест на 2023. Метрики усредняются. Для каждого окна вычисляются:

- AUC-ROC – площадь под ROC-кривой;

- Average Precision (AP) – усредненная точность по всем уровням полноты;

- Precision@20 – доля истинных новых связей среди 20 наиболее вероятных по версии модели.

Проверка статистической значимости различий между методами выполняется попарным тестом МакНемара ($\alpha = 0.05$).

Эксперименты планируется реализовать на языке Python с использованием NetworkX для работы с графами, Scikit-learn (NMF, метрики) и PyTorch (Dyngraph2vec, Node2Vec). Гиперпараметры обучаемых моделей (скорость обучения, количество эпох, размер батча) будут подбираться на валидационной части последнего года обучения. По завершении исследования код будет выложен в публичный репозиторий для воспроизводимости.

Список литературы

1. Goyal P. dyngraph2vec: Capturing network dynamics using dynamic graph representation learning / P. Goyal, S.R. Chhetri, A. Canedo // Knowledge-Based Systems. – 2020. – Vol. 187. – P. 104816. – DOI 10.1016/j.knosys.2019.06.024. – EDN XCJXHN.

2. Grover A. node2vec: Scalable feature learning for networks / A. Grover, J. Leskovec // Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD). – San Francisco, 2016. – P. 855–864.

3. Jin D. On generalizing static node embedding to dynamic settings / D. Jin, S. Kim, R.A. Rossi, D. Koutra // Proceedings of the 15th ACM International Conference on Web Search and Data Mining (WSDM). – Tempe, 2022. – P. 410–420.

4. Kumar A. Link prediction techniques, applications, and performance: A survey / A. Kumar, S.S. Singh, K. Singh, B. Biswas // Physica A: Statistical Mechanics and its Applications. – 2020. – Vol. 553. – P. 124289. – DOI 10.1016/j.physa.2020.124289. – EDN YFNJRN.

5. Lee D.D. Learning the parts of objects by non-negative matrix factorization / D.D. Lee, H.S. Seung // Nature. – 1999. – Vol. 401. – P. 788–791.

6. Newman M.E.J. Clustering and preferential attachment in growing networks / M.E.J. Newman // Physical Review E. – 2001. – Vol. 64, No. 2. – P. 025102.

7. Qin M. Temporal link prediction: A unified framework, taxonomy, and review / M. Qin, D.-Y. Yeung // ACM Computing Surveys. – 2023. – Vol. 56, No. 2. – P. 1–40.