УДК 004

Векторизация графов знаний с использованием моделей перехода на основе расстояния

Ветров Владислав Сергеевич – аспирант Финансового университета при правительстве РФ.

Аннотация: В современном мире одним из наиболее важных представлений структурированной информации является граф знаний. Его квантизация и исследование является одним из наиболее перспективных направлений в области анализа данных и развития искусственного интеллекта.

Для анализа и использования графа знаний в современных системах машинного обучения, его необходимо привести в удобную форму, например представить его в системе векторов. Многомерность вектора позволит заложить большое количество релевантной информации о различных свойствах графа в его элементы – сущности и отношения.

Ключевые слова: граф знаний, вектор, матрица, оптимизация, гиперплоскость, проекция.

Введение

Графы знаний (далее ГЗ) стали одним из ключевых ресурсов для развития машинного обучения и искусственного интеллекта, включая такие задачи, как создание вопросно-ответных систем, систем рекомендаций, поиска и создания знания и т. д. В последние годы появилось несколько крупных ГЗ, таких как Freebase и DBpedia, NELL, Wikidata и др. Данные графы были созданы путем машинного извлечения структурированной информации из различной текстовой информации в интернете и ручного обогащения этих данных.

Несмотря на то, что крупный ГЗ содержат миллиарды фактоидов-троек, структурированные знания до сих пор составляют малый процент от «абсолютного» знания и, вероятно, содержат множество различных ошибок и противоречий.

Из этого следует, что выявление новых связей в графе знаний, или даже его максимально-возможное завершение является одним из ключевых вопросов в исследовании ГЗ. Для достижения этого требуется процедура прогнозирования недостающей структурированной информации (сущностей и отношений) на основе существующих ГЗ.

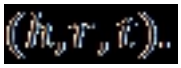
Типичный граф знаний структурирует реальную и абстрактную информацию в формате фактоидов-троек и обозначается как (головная сущность, отношение, хвостовая сущность), для краткости $(head, relation, tail)$ или (h, r, t). В формате графа сущности ранее упомянутые сущности являются вершинами, а отношения – ребрами.

Среди задач по восстановлению графов встречаются такие как, прогнозирование хвостовой сущности при наличии головной и отношения - $(h, r, ?)$, прогнозирование отношения - $(h, ?, t)$, и, наконец, прогнозирования головной сущности - $(?, r, t)$.

В процессе решения задач был создан целый класс успешных моделей перехода на основе расстояния, включая TransE, TransH, TransR, TransT и т. д. Эти модели направлены на создание наиболее точных векторных представлений для сущностей и отношений по принципу $h + r \approx t$, то есть t переходит в h с помощью r.

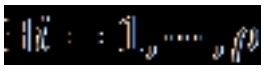
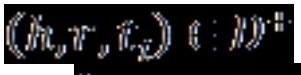
Обзор основных алгоритмов векторизации ГЗ с помощью перехода на основе расстояния

TransE является наиболее наглядной моделью в классе моделей на основе дистанции.

Данная модель представляет сущности и отношения как векторы в одном и том же пространстве, скажем 
. Если, к примеру, рассматривать фактоид (
– головная сущность, 
- отношение,
– хвостовая сущность), то отношение интерпретируется как вектор перехода от 
к
, такой что удовлетворяет следующему условию: отношение 
выполняется с низкой ошибкой при условии  .

TransE определяется как отрицательное расстояние между 
и
, т.е.

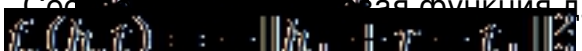
Несмотря на свою простоту и эффективность, TransE имеет недостатки при работе с отношениями по типу «1 к N», «N к 1» и «N к N». К примеру, можно рассмотреть отношение типа «1 к N». Если брать в рассмотрение отношение r,

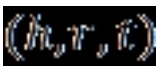

такое, что выполняется 
, TransE форсирует условие 
для всех 
, и следовательно 
. Это означает, что, например, при использовании TransE для векторизации двух троек (Эльдар Рязанов, режиссер, Служебный роман) и (Эльдар Рязанов, режиссер, Ирония судьбы, или С легким паром!) по условиям TransE должны будут выполняться два

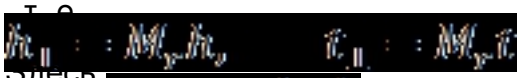
условия: Эльдар Рязанов + режиссер = Служебный роман и Эльдар Рязанов + режиссер = Ирония судьбы, или С легким паром!. Из этого можно вывести, что «Ирония судьбы, или С легким паром!» и «Служебный роман» являются одной и той же сущностью в векторном представлении TransE. Аналогичные недостатки существуют для отношений типа «N к 1» и «N к N».

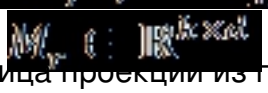
Рисунок 1. Иллюстрация работы алгоритма TransE.

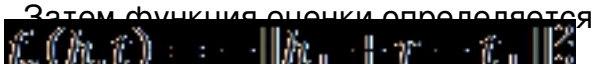
Такая модель как TransH продолжает и модернизирует идеи TransE. Для решения вышеперечисленных проблем модель вводит понятие гиперплоскостей в разрезе отношений, не сущностей. TransH моделирует объекты снова как векторы, но каждое отношение r как вектор **r** на гиперплоскости, где w_r - вектор нормали. Беря в рассмотрение фактоид (h, r, t), представления сущностей h и t являются сначала проецируется на гиперплоскость:

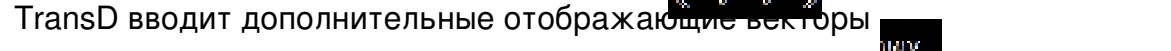


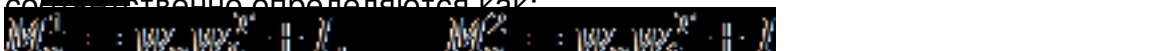


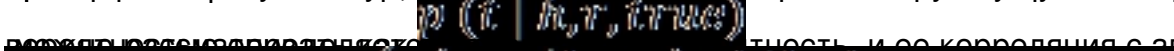


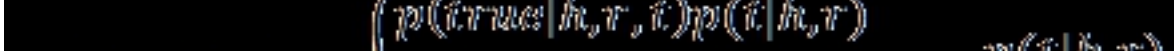


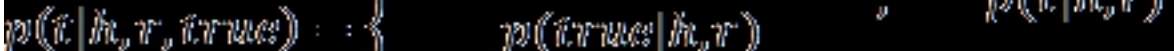


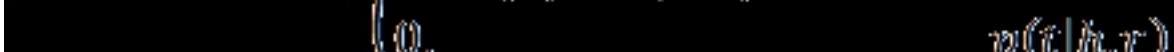












$$h_{\perp} = h - w_r^{\top}, \quad t_{\perp} = t - w_r^{\top} t w_r$$

[illegible]нены с малой ошибкой с помощью **r** на

т.е. $h_{\perp} + r \approx t_{\perp}$

[illegible] для оптимизации может быть представлена как:

$$f_r(h,t) = -\|h_{\perp} + r - t_{\perp}\|_2^2$$

[illegible] $\mathbb{R}^d$

, каждое отношение связано с определенным пространством $\mathbb{R}^k$

и моделируется как вектор перехода в этом пространстве. Если существует фактоид (h,r,t)

, TransR проецирует представления сущностей h

и t

в пространство, специфичное для отношения r

т.е.

$$h_{\perp} = M_r h, \quad t_{\perp} = M_r t$$

Здесь $M_r \in \mathbb{R}^{k \times d}$

- матрица проекции из пространства сущностей в пространство отношений r

Затем функция оценки определяется как,

$$f_r(h,t) = -\|h_{\perp} + r - t_{\perp}\|_2^2$$

[illegible] матрицы проекции в произведение двух (h,r,t)

TransD вводит дополнительные отображающие векторы w_h

, $w_t \in \mathbb{R}^d$ и $w_r \in \mathbb{R}^k$

вместе с представления сущностей / отношений $h, t \in \mathbb{R}^d$ и $r \in \mathbb{R}^k$

. Две матрицы проекции M_r^1

и M_r^2

соответственно определяются как:

$$M_r^1 = w_r w_h^{\top} + I, \quad M_r^2 = w_r w_t^{\top} + I$$

Затем эти две проекционные матрицы наносятся на голову объект h

и хвостовой объект t

соответственно, чтобы получить их проекции, т.е.

$$h_{\perp} = M_r^1 h, \quad t_{\perp} = M_r^2 t$$

[illegible] задачу $p(t \mid h, r, true)$

[illegible] и ее корреляция с априорной

$$p(t \mid h, r, true) = \begin{cases} \dfrac{p(true \mid h, r, t)\, p(t \mid h, r)}{p(true \mid h, r)}, & p(t \mid h, r) \neq 0 \\ 0, & p(t \mid h, r) = 0 \end{cases}$$

Заключение

Такие модели, как TransE, и от нее производные, позволяют получать точные векторные представления сущностей и отношений графа знаний. Несмотря на то, что базовая модель TransE обладает рядом недостатков, она подтолкнула к развитию целое семейство алгоритмов перехода на основе расстояния.

Данные модели предлагают не только различные подходы к векторизации графов знаний, но и возможности для добавления новой информации, например, типов. Не смотря на то, что у некоторых моделей могут быть сложности с обработкой определенных кейсов, например отношений по типу «1 к N» у TransE, или высокая алгоритмическая сложность у TransH, данные методы являются одними из наиболее интуитивных и точных алгортимов для векторизации, доступных на сегодня.

Список литературы

1. С. Николенко, А. Кадурин, Е. Архангельская, Глубокое обучение. – СПб.: Питер, 2018. – 281 стр.
2. Hongyun Cai, Vincent W. Zheng, and Kevin Chen-Chuan Chang, A Comphensive Survey of Graph Embedding: Problems, Techniques and Applications, arXiv: 1709.07604.Конец формы
3. Quan Wang, Zhendong Mao, Bin Wang, and Li Guo, Knowledge Graph Embedding: A Survey of Approaches and Applications, IEEE Transactions on Knowledge and Data Engineering (Volume: 29, Issue: 12, Dec. 1 2017), pp 2724 – 2743.
4. Shiheng Ma, Jianhui Ding, Weijia Jia, Kun Wang, and Minyi Guo, TransT: Type-Based Multiple Embedding Repsentations for Knowledge Graph Completion, ECML PKDD 2017: Machine Learning and Knowledge Discovery in Databases, pp 717-733.

{social}