УДК 004.021

ИССЛЕДОВАНИЕ МОДЕЛЕЙ СО СКРЫТЫМ РАССТОЯНИЕМ ДЛЯ ВНЕДРЕНИЯ СУЩНОСТЕЙ И ИХ ОТНОШЕНИЙ В ГРАФ ЗНАНИЙ

Хлопенкова А.Ю., Белов Ю.С.

Московский государственный технический университет имени Н.Э. Баумана, филиал, Калуга, e-mail: iu4-kf@mail.ru

В данной статье рассматриваются алгоритмы моделирования графа знаний, основанные на моделях со скрытым расстоянием. Демонстрируется представление внедряемых сущностей в векторном пространстве. Описывается роль использования триплетов, содержащих сущности и отношения, для построения графа. Для каждого из алгоритмов приводится функция вычисления обучаемой модели. Модель TransE представляет собой первоначальную реализацию с использованием триплетов (h, r, t) для представления отношений в виде трансляций в векторном пространстве и является базовой для всех остальных. В моделях TransH и TransR для реализации графа знаний подчеркивается необходимость использования гиперплоскостей. Так, например, отношение моделируется как вектор на собственной гиперплоскости. На примере модели TransD демонстрируется использование матрицы динамического отображения для внедрения сущностей. При обучении модели внедрению именованных символьных объектов (сущностей или отношений) TransD может учитывать их разнообразие. В модели KG2E приводится пример с использованием распределения Гаусса. В качестве меры расстояния используются расстояние Кульбака – Лейблера и оценочная вероятность. В заключение делаются выводы по каждой из представленных моделей. Подчеркиваются преимущества каждой из описанных моделей.

Ключевые слова: граф знаний, модели со скрытым расстоянием, триплеты, сущности, отношения, внедрение

RESEARCH OF LATENT DISTANCE MODELS FOR EMBEDDING OF ENTITIES AND THEIR RELATIONSHIPS IN THE KNOWLEDGE GRAPH

Khlopenkova A.Yu., Belov Yu.S.

Bauman Moscow State Technical University, Kaluga branch, Kaluga, e-mail: iu4-kf@mail.ru

The article discusses knowledge graph modeling algorithms based on latent distance models. The representation of embedded entities in vector space is demonstrated. The role of using triplets containing entities and relationships for graph construction is described. For each of the algorithms, the calculation function for the training model is given. The TransE model is an initial implementation which use triplets (h, r, t) to represent relationships in the form of translations in a vector space and it is basic for all others. TransH and TransR models emphasize the need for hyperplanes to implement the knowledge graph. So, for example, a relation is modeled as a vector on its own hyperplane. Using the TransD model as an example, the use of a dynamic mapping matrix for implementing entities is demonstrated. During the learning embeddings of named symbol objects (entities or relations), TransD could considers the diversity of them both. The KG2E model provides an example using the Gaussian distribution. The Kullback – Leibler distance and the estimated probability are used as a distance measure. Finally, draws conclusions for each of the presented models. The advantages of each described model are highlighted.

Keywords: knowledge graph, latent distance models, triplets, entities, relationships, embedding

Граф знаний – это мультиреляционный граф, состоящий из сущностей, которые могут быть представлены как в виде связки узлов и отношений, так и в виде разнотипных ребер. В качестве экземпляра ребра может быть представлена тройка параметров (голова, отношение, хвост), она обозначается как (h, r, t) от англ. (head, relation, tail).

Наиболее важные варианты использования графов знаний в объединении с ИИ включают в себя [1]:

– интуитивно понятный поиск с использованием естественного языка;

– обнаружение соответствующего контента и информации в структурированных или неструктурированных данных;

– надежный контент и управление данными;

– прогнозирование операционного риска; и т.п.

Цель исследования: рассмотреть модели со скрытым расстоянием, используемые для реализации графа знаний. Сравнить обучаемые модели и выделить особенности каждой из них.

Объект исследования

Одними из алгоритмов реализации графа знания являются модели со скрытым расстоянием. Данные модели используют функции рейтинговой оценки, основанные на расстоянии между сущностями, для моделирования тройки графа знаний. Взаимодействие сущностей и отношений осуществляется через их скрытое пространственное представление.

TransE. TransE (от англ. Translating Embeddings for Modeling Multi-relational Data – Транслирование внедряемых сущностей в векторном пространстве для моде-

лирования мультиреляционных данных) – это модель, основанная на EBM (от англ. Energy-Based Model), которая представляет отношения в виде трансляций в векторном пространстве. В частности, предполагается, что если триплет (h, r, t) имеет место, то внедрение сущности хвоста «t» должно быть близко к внедрению головной сущности «h» плюс некоторый вектор, который зависит от отношения «r». В TransE и сущности, и отношения являются векторами в одном и том же пространстве. Чтобы изучить такие внедрения, следует на обучаемой выборке минимизировать критерий ранжирования, основанный на границе Υ [2]:

$$\mathcal{L} = \sum_{(h,r,t)\in S(h',r',t')\in S'_{(h,r,t)}} \sum \left[\gamma + d\left(h+r,t\right) - d\left(h'+r,t'\right)\right]_+ ,$$

$$S'_{(h,r,t)} = \left\{\left(h',r,t\right)|h' \in E\right\} \cup \left\{\left(h,r,t'\right)|t' \in E\right\},$$

$$d\left(h+r,t\right) = \left|\left|h\right|\right|_2^2 + \left|\left|r\right|\right|_2^2 + \left|\left|t\right|\right|_2^2 - 2\left(h^T t + r^t\left(t-h\right)\right),$$

$$\left|\left|h\right|\right|_2^2 = \left|\left|t\right|\right|_2^2 = 1,$$

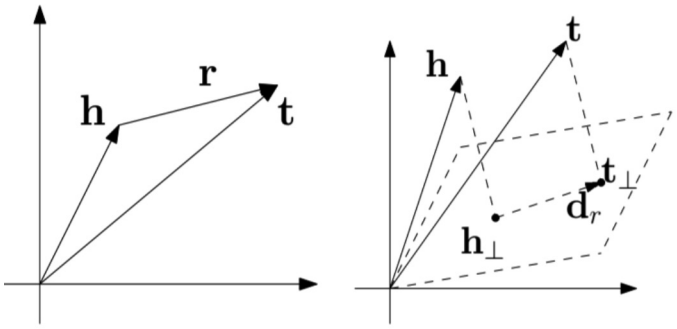
где $[]_+$ – представляет собой только положительную часть,
Υ – граница, разделяющая положительные и отрицательные тройки,
d – функция расстояния,
S – множество положительных триплетов (h, r, t),
S' – множество отрицательных триплетов,
$\mathcal{L}$ – обучаемая модель,
E – множество сущностей, $h, t \in E$.

TransH. TransH (от англ. Knowledge Graph Embedding by Translating on Hyperplanes – внедрение сущностей графа знаний путем транслирования на гиперплоскости) следует общему принципу TransE. Однако, по сравнению с ним, он вводит конкретные для отношений гиперплоскости. Сущности представлены в виде векторов, как и в TransE, однако отношение моделируется как вектор на собственной гиперплоскости с вектором нормали (рис. 1). Затем сущности проецируются на гиперплоскость отношения для расчета потерь. Чтобы стимулировать различие между хорошими триплетами и неправильными триплетами, используется следующий предельный рейтинг потери [3]:

$$\mathcal{L} = \sum_{(h,r,t)\in S(h',r',t')\in S'_{(h,r,t)}} \sum \left[\gamma + f_r(h,t) - f_r(h',t')\right]_+ ,$$

$$f_r\left(h,t\right) = \left(h - w_r^T h w_r\right) + d_r - \left(t - w_r^T t w_r\right)_2^2,$$

где w_r – гиперплоскость для отношения r,
f_r – функция оценки.



а) TransE *б) TransH*

Рис. 1. Иллюстрация алгоритмов моделирования графа знаний

TransR. TransR (от англ. Relation Embeddings for Knowledge Graph Completion – внедрение отношений для заполнения графа знаний) очень похож на TransH, с той лишь разницей, что вместо того, чтобы иметь одну гиперплоскость отношения, он вводит множество гиперплоскостей (рис. 2). Сущности являются векторами в пространстве сущностей, и каждое отношение является вектором в определенном пространстве отношения. Для расчета потерь сущности проецируются в отношении конкретного пространства с использованием матрицы проекции. Для каждой тройки (h, r, t) сущности в пространстве сущностей сначала проецируются в пространство r-отношений h_r и t_r с помощью операции Mr, а затем приводят к уравнению вида $h_r + r \approx t_r$ [4].

$$\mathcal{L} = \sum_{(h,r,t)\in S(h',r',t')\in S'_{(h,r,t)}} \max(0, \gamma + f_r(h,t) - f_r(h',t')),$$

$$f_r(h,t) = h_r + r - t_{r2}^2,$$

$$h_r = hM_r, t_r = tM_r,$$

TransD. TransD (от англ. Knowledge Graph Embedding via Dynamic Mapping Matrix – внедрение сущностей графа знаний с помощью матрицы динамического отображения) – улучшенная версия TransR. Для каждого триплета (h, r, t) используются две матрицы отображения M_{rh}, $M_{rt} \in R^{mn}$ для проецирования сущностей из пространства сущностей в пространство отношений. M_{rh}, M_{rt} сопоставляют матрицы h, t соответственно (рис. 3) [5].

$$\mathcal{L} = \sum_{\xi\in S}\sum_{\xi'\in S'} \left[\gamma + f_r(\xi') - f_r(\xi)\right]_+,$$

$$\xi = (h,r,t), \xi' = (h',r,t\,\hat{}\,'),$$

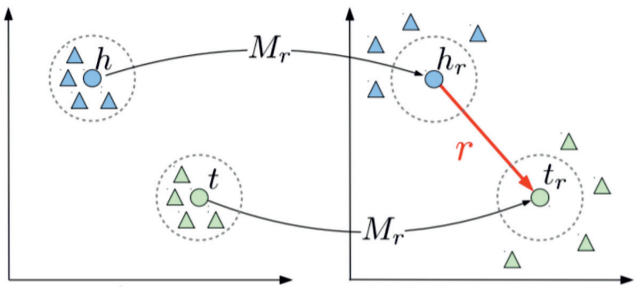
$$f_r(h,t) = -h_\perp + r - t_{\perp 2}^2,$$

$$h_\perp = M_{rh}h, t_\perp = M_{rt}t,$$

$$M_{rh} = r_p h_{ip} + I^{m*n},$$

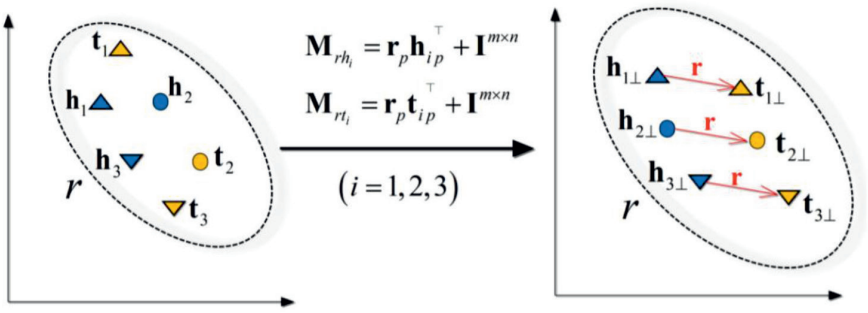
$$M_{rt} = r_p t_{ip} + I^{m*n},$$

где p – вектор проекции,
M_{rh}, M_{rt} – матрицы динамического отображения,
I^{m*n} – единичная матрица размером m×n.



Пространство сущностей *Пространство отношений*

Рис. 2. Иллюстрация алгоритма TransR

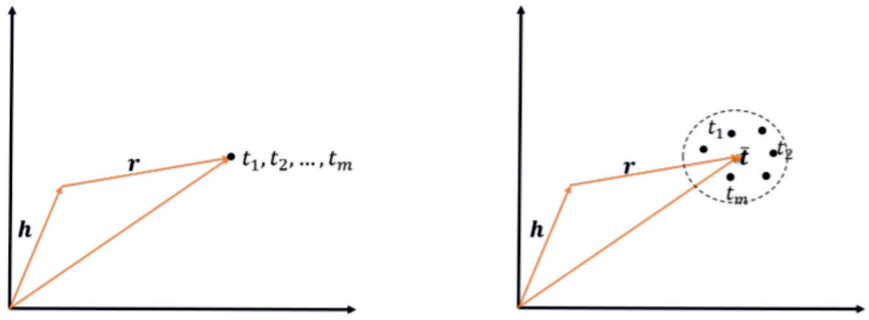


Пространство сущностей *Пространство отношений*

Рис. 3. Иллюстрация алгоритма TransD

TransM. TransM (от англ. Transition-based Knowledge Graph Embedding with Relational Mapping Properties – внедрение отношений для построения графа знаний на основе переходов со свойствами реляционного отображения) помогает устранить отсутствие гибкости алгоритма в TransE, когда дело доходит до сопоставления свойств триплетов. На рис. 4 продемонстрирована разница между алгоритмами. Используется структура графа знаний посредством предварительного вычисления индивидуального веса для каждого тестового триплета в соответствии с его свойством реляционного отображения. Простой способ смоделировать структуру – связать каждый тестовый триплет с весом, который представляет степень отображения. Согласно наблюдениям, свойство отображения триплета во многом зависит от его отношений внутри графа [6].

$$\mathcal{L} = \min \sum_{(h,r,t)\in S(h',r',t')\in S'_{(h,r,t)}} \left[\gamma + f_r(h,t) - f_r(h',t')\right]_+ .$$



а) TransE *б) TransM*

Рис. 4. Различия между алгоритмами при моделировании экземпляров отношения один ко многим

KG2E. KG2E (от англ. Knowledge Graphs with Gaussian Embedding – графы знаний с внедрением распределения Гаусса) вместо того, чтобы представлять сущности и отношения в качестве детерминированных точек в векторном пространстве, моделирует сущности и отношения (h, r, t), используя случайные величины, полученные из многомерного Гауссовского распределения. Затем KG2E оценивает события, используя отношение трансляций, оценивая расстояние между двумя распределениями r и t-h. KG2E предоставляет на выбор две меры расстояния (расстояние Кульбака – Лейблера и оценочная вероятность) [7].

$$\mathcal{L} = \sum_{(h,r,t)\in S(h',r',t')\in S'_{(h,r,t)}} \left[\gamma + \mathcal{E}(h,r,t) - \mathcal{E}(h',r',t')\right]_+ .$$

Расстояние Кульбака – Лейблера

$$\mathcal{E}(h,r,t) = \mathcal{E}(P_e P_r) = D_{KL}(P_e P_r) = \int_{x\in R^{k_e}} \mathcal{N}(x;\mu_r,\sum\nolimits_r) \log \frac{\mathcal{N}\left(x;\mu_e,\sum_e\right)}{\mathcal{N}\left(x;\mu_r,\sum_r\right)} dx) =$$

$$= \frac{1}{2}\left\{ tr\left(\sum\nolimits_r^{-1}\sum\right)_e + (\mu_r - \mu_e)^T \sum\nolimits_r^{-1} (\mu_r - \mu_e) - \log \frac{\det\left(\sum_e\right)}{\det\left(\sum_r\right)} - k_e \right\}.$$

Оценочная вероятность

$$\mathcal{E}(h,r,t) = \log \mathcal{E}(P_e P_r) = \log \mathcal{N}\left(0;\mu_e - \mu_r, \sum\nolimits_e + \sum\nolimits_r\right) =$$

$$= \frac{1}{2}\left\{ (\mu_e - \mu_r)^T \left(\sum\nolimits_e + \sum\nolimits_r\right)^{-1} (\mu_e - \mu_r) + \log\det\left(\sum\nolimits_e + \sum\nolimits_r\right) + k_e \log(2\pi) \right\},$$

$$\mathcal{E}\left(P_e P_r\right) = \int_{x \in R^{k_e}} \mathcal{N}\left(x; \mu_e, \sum_e\right) \mathcal{N}\left(x; \mu_r, \sum_r\right) dx\Big) = \mathcal{N}\left(0; \mu_e - \mu_r, \sum_e + \sum_r\right),$$

где $tr(\sum_r)$ – трассировка ковариационной матрицы,
$\mathcal{E}$ – множество сущностей триплета,
D_{KL} – расстояние Кульбака – Лейблера,
$\sum_e$, $\sum_r$ – ковариационные матрицы,
μ – вектор средних,
$\mathcal{N}\left(\mu, \sum\right)$ – многомерное Гауссово распределение (с диагональной ковариацией для эффективности вычислений),
P_e, P_r – распределение вероятностей и отношений соответственно,
k_e – размер сущности в скрытом векторном пространстве,
det – определитель матрицы.

Заключение

В заключение можно выделить основные характеристики описанных алгоритмов реализации графа знаний. При максимуме внимания на минимальной параметризации модели был разработан алгоритм TransE, чтобы в первую очередь представлять иерархические отношения. Он является хорошо масштабируемой моделью. TransH преодолевает недостатки TransE, связанные с рефлексивными отношениями «один ко многим / многие к одному / многие ко многим», при этом наследуя его эффективность. Обширные эксперименты классификации триплетов и извлечения реляционных сущностей показывают, что TransH приносит многообещающие улучшения в TransE. TransR предлагает внедрения посредством трансляции между проецируемыми сущностями. TransD имеет меньшую сложность и большую гибкость, чем TransR / CTransR. Обширные эксперименты показывают, что TransD превосходит TrasnE, TransH и TransR / CTransR по двум задачам, а именно классификации триплетов и прогнозированию ссылок. TransM – превосходная модель, которая не только идеальна для представления иерархических и нерефлексивных характеристик, но также гибкая для адаптации различных свойств отображения триплетов знаний. Результаты обширных экспериментов с несколькими эталонными наборами данных доказывают, что модель может достичь более высокой производительности без ущерба для эффективности. KG2E представляет собой новый метод для представлений сущностей и отношений. Обширные эксперименты по прогнозированию ссылок и классификации триплетов с несколькими наборами эталонных данных (включая WordNet и Freebase) демонстрируют, что предлагаемый метод превосходит аналоговые методы.

Список литературы

1. Хлопенкова А.Ю., Белов Ю.С. Методы обработки естественного языка в виртуальных голосовых помощниках // E-Scio Электронное периодическое издание «E-Scio.ru» – Эл № ФС77-66730. [Электронный ресурс]. URL: http://e-scio.ru/wp-content/uploads/2019/11/Хлопенкова-А.-Ю.-Белов-Ю.-С..pdf (дата обращения: 01.07.2020).

2. Antoine Bordes, Nicolas Usunier, Alberto Garcia-Duran, Jason Weston, Oksana Yakhnenko. Translating Embeddings for Modeling Multi-relational Data. Advances in Neural Information Processing Systems 26. 2013. [Electronic resource]. URL: http://papers.nips.cc/paper/5071-translating-embeddings-for-modeling-multi-relational-data.pdf (date of access: 01.07.2020).

3. Zhen Wang, Jianwen Zhang, Jianlin Feng, Zheng Chen. Knowledge Graph Embedding by Translating on Hyperplanes. AAAI Publications, Twenty-Eighth AAAI Conference on Artificial Intelligence. 2014. [Electronic resource]. URL: https://persagen.com/files/misc/wang2014knowledge.pdf (date of access: 01.07.2020).

4. Yankai Lin, Zhiyuan Liu, Maosong Sun, Yang Liu, Xuan Zhu. Learning Entity and Relation Embeddings for Knowledge Graph Completion. Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence. 2015. [Electronic resource]. URL: http://nlp.csai.tsinghua.edu.cn/~lyk/publications/aaai2015_transr.pdf (date of access: 01.07.2020).

5. Guoliang Ji, Shizhu He, Liheng Xu, Kang Liu, Jun Zhao. Knowledge Graph Embedding via Dynamic Mapping Matrix. Association for Computational Linguistics. 2015. DOI 10.3115/v1/P15-1067. [Electronic resource]. URL: https://www.aclweb.org/anthology/P15-1067.pdf (date of access: 01.07.2020).

6. Miao Fan, Qiang Zhou, Emily Chang, Thomas Fang Zheng. Transition-based Knowledge Graph Embedding with Relational Mapping Properties. Department of Linguistics, Chulalongkorn University. 2014. [Electronic resource]. URL: https://www.aclweb.org/anthology/Y14-1039.pdf (date of access: 01.07.2020).

7. Shizhu He, Kang Liu, Guoliang Ji and Jun Zhao. Learning to Represent Knowledge Graphs with Gaussian Embedding. Conference: the 24th ACM International. 2015. DOI: 10.1145/2806416.2806502. [Electronic resource]. URL: http://www.nlpr.ia.ac.cn/cip/~liukang/liukangPageFile/Learning%20to%20Represent%20Knowledge%20Graphs%20with%20Gaussian%20Embedding.pdf (date of access: 01.07.2020).