Обзор основных методов автоматического обновления и расширения графов знаний на основе новых данных с точки зрения корректных разрешений возможных конфликтов при слиянии данных.

Ключевые слова: графы знаний, расширение, дополнение, конфликты при слиянии данных.

Аннотация

Данная статья посвящена проблеме обновления и расширения графов знаний с точки зрения корректных разрешений возможных конфликтов при слиянии данных на фоне ежегодного увеличения объема генерируемой человечеством информации. В статье выделены ключевые критерии, такие как интерпретируемость выводов методов, вычислительная сложность и сложность построения метода, для сравнения шести основных методов дополнения графов знаний. На основе сравнений методов были выявлено, что существующие решения, обладая хорошей интерпретируемостью, сталкиваются с проблемами вычислительной сложности, в частности при масштабировании графов знаний, и сложности при построении самого метода. А решения, которые при масштабировании работают, в большинстве случаев имеют проблемы с интерпретируемостью полученных методом результатов.

Введение

Объем генерируемых человечеством данных с каждым годом увеличивается: только за последние два года (к 2023 году) было получено примерно 90% мировых данных, а на 2023 год генерация составила 120 зеттабайт [7]. Хранение и структурирование подобных данных в виде графов знаний используются во многих современных информационных технологиях, в частности, в системах автоматической обработки текстов (в том числе, в семантических системах) [8] или в рекомендательных системах, например, в рекомендательной системе для агрегирования междисциплинарных знаний относительно улучшения психического здоровья в графе эмоциональных знаний (EmoKG) [9] или фармакологической рекомендательной системе лекарственных средств [10]. С ростом количества получаемых данных возрастает и сложность автоматического обновления и расширения графов знаний с учетом корректного разрешения возможных конфликтов между новыми данными и между новыми и уже имеющимися данными, из-за чего используемые методы обновления и расширения графов знаний могут нуждаться в обновлении. Для того, чтобы понять, как с этим справляются существующие методы, были выделены критерии интерпретируемости выводов методов при дополнении графа знаний, сложности вычислений при дополнении графа знаний, а также сложности построения и обучения метода. По этим критериям сравнивались шесть основных методов, а выводы, получившиеся после сравнения методов показали, что нынешние решения при хорошей

интерпретируемости зачастую имеют проблемы с вычислительной сложностью при масштабировании графа знаний, а также проблемы с сложностью построения.

Обзор предметной области

Принцип отбора аналогов

Для поиска методов автоматического расширения графов знаний на основе новых данных использовалась система поиска “Академия Google”. Поиск осуществлялся по следующим запросам: <<knowledge graph expansion “knowledge graph expansion”>>, <<knowledge graph completion “knowledge graph completion”>>. “Knowledge graph completion” (KGC) - устаявшийся термин для описания проблемы заполнения графа новыми данными [1]. Все методы заполнения можно разделить на два вида: так называемые, традиционные методы заполнения графов знаний (traditional knowledge graph completion methods) и методы, основанные на глубоком обучении представлений (deep representation learning). #### Расширение графа знаний на основе логических правил (Rule Reasoning Model) - Традиционный метод Метод заполнения графа знаний на основе логического рассуждения использует правила или статистические характеристики для вывода новых знаний, расширяя структуру графа и дополняя его [2]. Поскольку правила автоматически генерируются в соответствии с семантикой или извлекаются вручную, преимущество метода заключается в его высокой интерпретируемости и точности создаваемых данных в графе знаний. В то же время этот метод также имеет недостатки. Прежде всего, этот метод сильно зависит от правил, которые построить вычислительно трудно, независимо от способа их построения (ручного или автоматического). Причем, при увеличении масштаба графа знаний вычисления новых правил возрастает в разы, из-за чего этот метод становится неприменим. #### Расширение графа знаний на основе вероятностной модели графа (Probabilistic Graph Model) - Традиционный метод Метод заполнения графа знаний на основе вероятностной графовой модели использует графы для представления вероятностных отношений, обеспечивая меньшую вычислительную сложность по сравнению с методами, основанными на правилах [3]. Этот метод преимущественно использует моделирование сетей Маркова и Байесовские сети. Сети Маркова используются для представления вероятностных связей, объединяя графовую структуру с теорией вероятностей. Байесовские сети учитывают структуру сети и информацию об атрибутах узлов. Они представляют собой направленный ациклический граф. Преимущества такого подхода: гибкая топологическая структура, простота интерпритируемости (процесс рассуждения в алгоритме объясним). В то же время она хорошо работает с точки зрения повышения точности прогнозирования и сокращения временных затрат. Однако из-за высокой сложности алгоритма его трудно рассчитать для масштабных графов знаний с несколькими отношениям. #### Расширение графа знаний на основе вычислений графа (Graph Calculation Model) - Традиционный метод Метод заполнения графа знаний на основе графового вычисления абстрагирует структуру графа знаний в виде графа, где сущности представлены узлами, а отношения различных типов действуют как рёбра [4]. С использованием статистических характеристик узлов и рёбер, таких как степень узла и матрица смежности, можно предсказывать новые сущности и

отношения. Этот метод легко поддается интерпретации и не требует дополнительных логических правил, помогающих в процессе рассуждения. Имеет проблемы с масштабируемостью, высокого использования памяти и сталкивается с проблемой сложности крупномасштабных вычислений данных. #### Расширение графа знаний на основе модели перевода (Translation Model) - Глубокое обучение представлений Расширение графа знаний, основанное на модели трансляции, предсказывает, что новые отношения сущностей утраиваются из существующего графа знаний путем встраивания сущностей и отношений в векторное пространство. Большинство существующих методов фокусируются на структурированной информации троек (субъект, отношение, объект) и максимизируют возможность их установления [5]. Метод заполнения графа знаний, основанный на модели перевода, фокусируется на использовании взаимосвязи между сущностями, семантики, содержащейся в сущности и отношениях, и структурированной информации графа знаний для реализации моделирования сущностей и отношений, что компенсирует сложное обучение и трудное расширение традиционных методов. Для моделирования сущностей и отношений методы очень просты и понятны с высокой интерпретируемостью #### Расширение графа знаний на основе модели семантического сопоставления (Semantic Matching Model) - Глубокое обучение представлений Данный метод основан на семантическом сходстве и использует функцию оценки, основанную на семантической схожести, для извлечения потенциальных семантических ассоциаций между сущностями и отношениями [6]. Путем вложения представлений сущностей и отношений в векторное пространство метод может предсказывать новые факты и дополнять граф знаний. Этот метод обеспечивает высокую точность при прогнозировании симметричных отношений, может хорошо выявлять потенциальные семантические ассоциации. В то же время существуют проблемы со многими параметрами модели и высокой вычислительной сложностью, которые не могут быть адаптированы для заполнения крупномасштабных графов знаний #### Расширение графа знаний на основе обучения представлений сети (Network Representation Learning Model) - Глубокое обучение представлений Метод, основанный на обучении сетевому представлению, направлен на объединение информации, извлеченной из структуры топологии сети, и информации о содержимом узлов и ребер, преобразование вершин сети во встраиваемые представления в низкоразмерном непрерывном векторном пространстве и реализацию задачи завершения графа знаний с помощью машинного обучения. Применение этого метода, основанного на сетевом представлении, к задаче завершения графа знаний позволяет лучше извлекать скрытые функции в структуре графа знаний, что полезно для обучения модели прогнозирования. Все вышеупомянутые сети являются неглубокими сетями, и трудно иметь дело с сильно нелинейными и разреженными сетевыми структурами. ### Критерии сравнения аналогов

Интерпретируемость

Для автоматического расширения графа знаний на основе новых данных важно понимание того, какие знания были добавлены и какие выводы сделаны.

Сложность вычислений

Методы, способные быстрее обрабатывать большие объемы данных для расширения графа, будут более подходящими для автоматизированных систем.

Сложность построения и обучения метода

Автоматическое расширение графа знаний требует методов, которые могут быть легко настроены и обучены на новых данных.

Таблица сравнения по критериям

Сравнение по критериям представлено в табл. 1.

Таблица 1 – Сравнение

	Интерпретируемость	**Сложность вычислений**	**Сложность построения и обучения модели**
Rule Reasoning Model	высокая интерпретируемость, так как составляемые правила формулируются на основе логической структуры знаний и отношений в графе (то есть можно проследить, какие шаги и условия привели к конкретному выводу)	проблемы с вычислительной сложностью при увеличении масштаба графа знаний	сложность построения правил может быть высокой при автоматическом создании правил
Probabilistic Graph Model	высокая интерпретируемостью (хоть и более низкая по сравнению с Rule Reasoning Model) благодаря использованию графов для представления вероятностных отношений	уменьшает сложность вычислений по сравнению с методами, основанными на правилах, но может столкнуться с проблемами для крупномасштабных графов	требует разработки сложных алгоритмов, основанных на моделях сетей Маркова и Байесовских сетях

	Интерпретируемость	Сложность вычислений	Сложность построения и обучения модели
	(вероятностные выводы основываются на известных графовых вероятностных моделях)		
Graph Calculation Model	высокая интерпретируемость, поскольку использует графовую структуру для предсказания новых сущностей и отношений	имеет проблемы с масштабируемостью и высоким использованием памяти при обработке крупных графов	требует статистических характеристик узлов и рёбер, но не требует дополнительных логических правил
Translation Model	для моделирования сущностей и отношений высоко интерпретируемы (в основе метода лежит геометрическая интерпретация векторных пространств - сущности и отношения представляются в виде векторов в пространстве, а операции над векторами используются для представления отношений между сущностями). Интер претируемость модели, как правило, хуже, чем у традиционных методов, однако из	может сталкиваться с вычислительной сложностью из-за применения глубокого обучения и обработки больших объемов данных	требует обширного обучающего набора данных и вычислительных ресурсов для обучения глубоких моделей

	Интерпретируемость	Сложность вычислений	Сложность построения и обучения модели
	нетрадиционных методов - это наиболее интерпретируемая модель		
Semantic Matching Model	обычно менее интерпретируемы, так как вовлекают сложные математические модели	высокая вычислительная сложность, которая не может быть адаптирована для заполнения крупномасштабных графов знаний	требует обширного обучающего набора данных и вычислительных ресурсов для обучения глубоких моделей
Network Representation Learning Model	обычно менее интерпретируемы, так как вовлекают сложные математические модели	может сталкиваться с вычислительной сложностью из-за применения глубокого обучения и обработки больших объемов данных	требует обширного обучающего набора данных и вычислительных ресурсов для обучения глубоких моделей

Выводы по итогам сравнения

Выбор метода для автоматического расширения графа знаний зависит от конкретных требований задачи, доступных ресурсов и важности интерпретируемости результатов. Каждый метод имеет свои преимущества и ограничения, и выбор должен быть обоснован учетом специфики задачи и возможностей инфраструктуры. В контексте задачи дополнения графов знаний новыми данными, где возможны конфликты при слиянии и необходимости их обрабатывать, наиболее важным критерием становится интерпретируемость выводов выбранного метода, что в большей мере свойственно традиционным методам. Тем не менее, существующие традиционные методы сталкиваются с проблемой масштабируемости графов знаний, что на фоне ежегодного увеличения объема генерируемых человечеством данных [7] становится существенной проблемой.

Выбор метода решения

Метод решения задачи дополнения графов знаний новыми данными, где возможны конфликты при слиянии и необходимости их обрабатывать, должен достигать такой же высокой интерпретируемости, так как как и традиционные методы, так как важно понимание того, какие знания были добавлены и какие выводы сделаны. Но при этом быть более пригодным для крупномасштабных графов, потому как с каждым годом увеличивается объем генерируемых человечеством данных, с чем традиционные модели справляются плохо.

Описание метода решения

На данный момент существует два наболее подходящих варианта решения задачи дополнения графов знаний новыми данными, где возможны конфликты при слиянии и необходимости их обрабатывать: либо использовать один из традиционных методов, потому как они обладают высокой степенью интерпретируемости, но низкой степенью масштабируемости, либо же использовать метод Translation Model, который хоть и не обладает той же высокой степенью интерпретируемости, однако лучше работает на крупномасштабных графах.

Заключение

Обзор основных методов дополнения графов на основе новых данных с точки зрения корректных разрешений возможных конфликтов при слиянии данных показал, что существующие решения могут справляться с данной задачей на небольших данных, но при масштабировании графа знаний все традиционные методы сталкиваются с проблемой вычислительной сложности из-за специфики построения метода. Методы, которые лучше работают при масштабировании, напротив же, зачастую имеют проблемы с интерпретируемостью, так как вовлекают более сложные математические модели.

В силу того, что один из методов (Translation Model) продемонстрировал, что добиться компромисса между интерпретируемостью и масштабируемостью возможно, в дальнейшем планируется создание метода, основанного на традиционных методах, который бы показывал более лучшие результаты при масштабировании графа знаний.

Список литературы

[1] Chen Z. et al. Knowledge graph completion: A review //Ieee Access. – 2020. – Т. 8. – С. 192435-192456.

[2] Carlson A. et al. Toward an architecture for never-ending language learning //Proceedings of the AAAI conference on artificial intelligence. – 2010. – Т. 24. – №. 1. – С. 1306-1313.

[3] Jiang S., Lowd D., Dou D. Learning to refine an automatically extracted knowledge base using markov logic //2012 IEEE 12th International Conference on Data Mining. – IEEE, 2012. –

C. 912-917.

[4] Lao N., Cohen W. W. Relational retrieval using a combination of path-constrained random walks //Machine learning. – 2010. – T. 81. – C. 53-67.

[5] Wang Q. et al. Knowledge graph embedding: A survey of approaches and applications //IEEE Transactions on Knowledge and Data Engineering. – 2017. – T. 29. – №. 12. – C. 2724-2743.

[6] Yang B. et al. Embedding entities and relations for learning and inference in knowledge bases //arXiv preprint arXiv:1412.6575. – 2014.

[7] Hassani H., MacFeely S. Driving Excellence in Official Statistics: Unleashing the Potential of Comprehensive Digital Data Governance //Big Data and Cognitive Computing. – 2023. – T. 7. – №. 3. – C. 134.

[8] Schneider P. et al. A decade of knowledge graphs in natural language processing: A survey //arXiv preprint arXiv:2210.00105. – 2022.

[9] Gyrard A., Boudaoud K. Interdisciplinary iot and emotion knowledge graph-based recommendation system to boost mental health //Applied Sciences. – 2022. – T. 12. – №. 19. – C. 9712.

[10] Sosa D. N. et al. A literature-based knowledge graph embedding method for identifying drug repurposing opportunities in rare diseases //Pacific Symposium on Biocomputing 2020. – 2019. – C. 463-474.