DOI 10.37882/2223–2966.2022.11–2.01

МЕТОДЫ ПРЕДСТАВЛЕНИЯ ТЕКСТОВЫХ ДОКУМЕНТОВ НА ОСНОВЕ ГРАФОВ В ЗАДАЧАХ ОБРАБОТКИ ЕСТЕСТВЕННОГО ЯЗЫКА

METHODS FOR GRAPH-BASED REPRESENTATION OF TEXTUAL DOCUMENTS IN NATURAL LANGUAGE PROCESSING TASKS

E. Akbasheva
G. Akbasheva
I. Tlupov

Summary. Nowadays, text is the most common form of information storage. Document representation is an important step in the process of data and text mining, natural language processing and information retrieval.

The paper provides a formal description of the graph-based text representation problem, also discusses some natural language processing (NLP) problems which depend on different types of graph structures, advantages and disadvantages of methods used for graph-based text representation. In addition, a comparison of different approaches is given, as well as the applications in which these approaches are used.

Keywords: graph, mathematical model, text analysis, natural language processing.

Акбашева Евгения Амировна
Старший преподаватель
Кабардино-Балкарский государственный университет
Нальчик
akbash_e@mail.ru
Акбашева Галина Амировна
Старший преподаватель
Кабардино-Балкарский государственный университет
Нальчик
galina_akbash@mail.ru
Тлупов Ислам Заурбекович
Кабардино-Балкарский государственный университет
Нальчик
tlupovislam@gmail.com

Аннотация. В настоящее время текст является наиболее распространенной формой хранения информации. Представление документа является важным этапом в процессе интеллектуального анализа данных и текста, обработки естественного языка и поиска информации.

В статье приводится формальное описание задачи представления текста на основе графа, также рассматриваются некоторые задачи обработки естественного языка (NLP), которые зависят от различных типов графовых структур, преимущества и недостатки методов, используемых для представления текста на основе графов. Кроме того, приводится сравнение различных подходов, а также областей применения, в которых используются эти подходы.

Ключевые слова: граф, математическая модель, текстовый анализ, обработка естественного языка.

Введение

В настоящее время огромное количество текста, генерируемого в различных областях человеческой жизни, делает необходимым создание и использование методов, позволяющих получить модели данных. Очевидно, что эффективный текстовый анализ в значительной степени зависит от способа представления структуры текстового документа.

В эпоху больших данных текст является одним из самых распространенных типов обработки. Представление данных является важным шагом в процессе извлечения элементов интеллектуального анализа данных. Таким образом, существует постоянная проблема в определении правильной модели представления текста, которая может в значительной степени отразить присущие текстовым данным особенности.

Следовательно, важной задачей является представлять структуру и семантическое содержание текстовых документов, а также связи между определенными частями документов, возникающие в процессе работы с ними. Одна из математических моделей представления текстов — это модель векторного пространства. Данная модель рассматривает числовые векторы признаков в евклидовом пространстве. Но модель векторного пространства не позволяет выразить смысл текста

и определить его структуру, слова рассматриваются независимо друг от друга, то есть последовательности слов не учитываются при анализе.

Слова организуются в разделы, параграфы, предложения и пункты для определения смысла документа. Следовательно, отношения между различными компонентами документа, их порядок и их расположение важны для детального понимания документа. Графовая модель представления текста известна как одно из лучших решений для этих проблем [1].

Графовое представление является математической конструкцией и может эффективно моделировать отношения и структурную информацию, а также может помочь в большинстве операций с текстом, таких как топологические, реляционные, статистические и т.д. В данной работе представлен обзор различных методов моделирования текстовых документов с помощью графов.

1. Документ как граф

Документ можно представить в виде графа [2], где термины представлены вершинами, а отношения между терминами — ребрами (1) — (2).

$$G = \{V, E\}. \quad (1)$$

В графовом представлении обычно существует 5 типов вершин: $V = \{F, S, P, D, C\}$, где F — термины, S — предложения, P — абзацы, D — документы, C — понятия.

$$F = \{t_1, t_2, \dots, t_n\}, S = \sum_{i=0}^{n} t_i, P = \sum_{i=0}^{n} s_i,$$

$$D = \sum_{i=0}^{n} p_i, DC = \sum_{i=0}^{n} d_i. \quad (2)$$

Отношения E между терминами могут быть синтаксические, статистические, семантические и могут отличаться в зависимости от контекста графа.

В традиционной модели представления текста «мешок слов» [3] информация о порядке или структуре слов в документе не учитывается. Модель предоставляет информацию только о том, встречаются ли известные слова в документе, а не об их положении, то есть она не позволяет определять важность терминов. В связи с этим возникает необходимость в другой модели представления текстовых документов. Графовое представление текста предлагается в качестве решения проблемы недостатков подхода «мешок слов».

Сети совпадений (или коокурренции, или совместной встречаемости) — одна из наиболее популярных моделей представления текста [4]. По сравнению с моделью «мешок слов» эта модель обеспечивает важный контекст для описания взаимосвязей между словами. Текст представляется в виде графа, где вершины отображают совпадения слов и ребер. Сети создаются путем соединения пар терминов с помощью набора критериев, определяющих их совпадение. Например, можно сказать, что термины A и B «совпадают», если они оба встречаются в определенном тексте. Другой текст может содержать термины B и C. Связывание A с B и B с C создает сеть совместной встречаемости этих трех терминов. Было обнаружено, что сети совместной встречаемости особенно полезны для анализа больших текстов и больших данных.

Новые модели получают высокую оценку из-за простоты и недостатков традиционных моделей. Текстовое представление, основанное на графах, может быть признано одним из реальных решений вышеперечисленных недостатков.

2. Представление текстового документа графом

Текстовый документ может быть представлен графом несколькими способами. При этом узлы обозначают характеристики, а ребра представляют отношения между узлами.

2.1. Граф коокурренции (совместной встречаемости)

Подходы к построению графа на основе совместной встречаемости слов в документе достаточно разнообразны.

Существует подход, при котором избегаются синтаксические фильтры, а отдельные признаки учитывается при построении графа [5], при этом если в тексте появляется новый термин, то в граф добавляется узел. Если они встречаются в пределах определенного размера окна, добавляется неориентированное ребро. Предложения связаны, если они находятся рядом или имеют общее слово.

Рассмотрим алгоритм, в котором предложение рассматривается как вершина, и предложения соединяются, если они находятся рядом друг с другом или имеют хотя бы одно общее ключевое слово [6]. Последовательные предложения в текстовом документе S_1, S_2, …, S_n представлены как множество вершин графа. Для каждого последовательного предложения (S_i, S_{i+1}) добавляется ребро. Тогда если два предложения имеют хотя бы одно общее слово, они могут соединяться с помощью ребра.

В спектральном методе [7] рассматривается синтаксическая связь между словами. Статистический метод [8] используется для поиска часто встречающихся слов.

Можно представить текстовый документ как взвешенный граф, в котором термины определены как узлы, ребро показывает связь между узлами в единице и вес, измеряющий силу связи. Минимальная длина предложения в качестве единицы измерения выбирается для измерения информации о совпадении терминов признаков вместо целого абзаца в качестве единицы измерения, чтобы избежать увеличения графа с потерей взаимной информации терминов признаков.

Для расчета силы связи используется формула (3):

$$W_{ij} = \frac{f(t_i,t_j)}{f(t_i)+f(t_j)-f(t_i,t_j)}, \quad (3)$$

где W_{ij} — это вес между вес между n_i и n_j, $f(t_i, t_j)$ — количество раз, когда t_i и t_j встречаются вместе в единице текста, $f(t_i)$ и $f(t_j)$ — частоты появления t_i и t_j в d_i соответственно. Высокие значения W_{ij} обозначают сильную связь, в противном случае — слабую.

Известный алгоритм для извлечения слов TextRank [9] извлекает репрезентативные слова из текстового документа. Эти слова представляются как вершины. Ненаправленные ребра между двумя вершинами вычисляются с помощью отношения взаимной встречаемости на основе расстояния между встречами слов, так что две вершины соединяются, если их соответствующие лексические единицы встречаются в окне максимального количества слов, которое может составлять от 2 до 10 слов.

2.2. Совместная встречаемость на основе POS-тегов

Целью POS-тегирования [10] является присвоение правильной лексической категории каждому слову в тексте. Основная трудность POS-тегирования заключается в том, что присвоение класса слову часто является неоднозначной задачей, поскольку лексическая категория слова обычно зависит от контекста, в котором оно используется. Например, одно и то же слово может быть использовано как в качестве существительного, так и в качестве глагола. Чтобы справиться с этой неоднозначностью, обычно рассматривают последовательности из n слов, чтобы вывести контекст, в котором слова используются. Эта альтернативная синтаксическая модель учитывает отношения между словами.

2.3. Семантический граф

Графовые модели обладают способностью отражать структурную информацию в текстах, но они не учитывают семантические отношения между словами. Для построения графа с учетом семантических отношений между словами используются граф тезаурусов и граф понятий [11]. В графе тезаурусов вершины обозначают термины, а ребра — смысловые отношения.

Граф понятий строится на основе текстового документа. Для поиска семантических ролей в предложении используются известные лексические языковые базы WordNet и VerbNet, и с помощью этих ролей строится граф понятий. «Сырой» текст предварительно обрабатывается, а недвусмысленные существительные сопоставляются с понятиями WordNet. Понятия, в отличие от слов, являются очень эффективным, лаконичным представлением содержания документа. Оно может быть легко и четко интерпретировано.

С помощью метода TF-IDF [12] можно оценить важность конкретного термина в контексте всего документа, входящего в выборку документов. Метрика TF/IDF является статистическим показателем и высчитывается как отношение между частотностью вхождения термина в текстовый документ и обратной частотой документа. Обратная частота документа — это инверсия частотности, с которой некоторое слово встречается в выборке текстов.

Показатель TF/IDF используют поисковые алгоритмы для определения релевантности текста в поисковых запросах. Также этот показатель можно использовать для определения близости документов друг другу, что может быть использовано при их группировке.

2.4. Анализ формальных понятий

Анализ формальных понятий — это основной метод, используемый для упорядочения объектов и свойств в понятийной иерархии или формальной онтологии. Каждое понятие представлено в иерархии как коллекция объектов, которые имеют общие сходные свойства для определенной группы свойств. Объекты в методе называются «формальными объектами» [4]. При поиске данных документация может рассматриваться как «объектоподобная», в то время как слова могут рассматриваться как «атрибутоподобные» [13]. Более того, такие элементы, как лексемы и виды вещей, качества и информация, составляют группу формальных элементов и их формальных качеств [14].

Анализ формальных понятий имеет практическое применение в области получения и интеллектуального анализа данных, получения контента, управления обучением, машинного обучения, разработки программ, исследований, семантического веба и т.д.

2.5. Концептуальный фреймовый граф

Концептуальный фреймовый граф представляет собой схему представления, подобную сценариям и ори-

Таблица 1. Сравнение различных схем представления текста графом

Модель	Достоинства	Недостатки
Граф совместной встречаемости	Высокий уровень производительности в стандартных задачах оценки	Размер окна совместной встречаемости
Совместная встречаемость на основе POS-тегов	Обобщение на основе модели приближается к человеческой производительности в плане поиска сходств между текстами	Необходимость использования внешнего POS-тега
Семантический граф	Эффективное, краткое представление содержания документа, которое можно легко и четко интерпретировать	Необходимость использования внешней онтологии
Модель концептуальных графов	Концептуальный граф универсален и доступен на различных уровнях деятельности по разработке информационной системы	Модель представляет собой неупорядоченный набор понятий и поддерживает только данные
Анализ формальных понятий	Способность обнаруживать взаимосвязь между терминами	Сложность в вычислительном времени по сравнению с другими графами
Концептуальный фреймовый граф	Обеспечивает описание, цели, семантику фреймов и семантическую роль для каждого термина в тексте	Неглубокая семантическая информация

ентированную на включение в строго организованные структуры данных неявных информационных связей, существующих в предметной области. Это представление поддерживает организацию знаний в более сложные единицы, которые отображают структуры объектов этой области.

В сценарии обмена знаниями интуитивная система описания понятий реализована на основе базы обучения. Данный метод является перспективным подходом для получения большего количества данных из достоверных документов и реальной жизни.

Для определения слов в тексте необходимо сначала определить этапы предварительной обработки, такие как стемминг — нахождение основы слова, леммы — начальные формы слова и т.д. С помощью алгоритма стемминга, или других методов, каждый термин в документе становится узлом в графе для нормализации языкового алгоритма. Все узлы в графе уникальны и различны, поскольку каждый узел имеет свой термин, даже если один и тот же термин повторяется в одном документе.

При выполнении извлечения данных с использованием графового представления текста и без него, алгоритмы, которые не были эффективны при использовании других графических подходов по сравнению с методом концептуального фреймового графа, получили улучшение точности и отзыва на 35% и 18% соответственно [15].

2.6. Модель концептуальных графов

Модель концептуальных графов представляет собой формализацию знаний, позволяющую моделировать семантику естественного языка [16]. Этот тип графов используется для извлечения текстовых признаков или работы по классификации языка для представления знаний [17, 18]. Этот подход хорошо известен в психологии, философии и лингвистике. Информационная структура на семантическом уровне может быть выражена в концептуальных графах. Концептуальные графы являются двудольными, связными и конечными [2]. Результирующая диаграмма содержит массив ребер и вертикальных узлов. Концептуальные графы различают отношения любой степени сложности и все, что остается в диалекте системы, с помощью кругового сегмента. Они могут рассматривать точно и глубоко организованные данные. Построенный концептуальный граф часто используется для графов планирования. Лингвистическая структура содержимого берется за основу для синтаксического анализа проектов перед преобразованием в концептуальный граф.

Концептуальные графы используются в различных областях. Например, в больницах они используются для получения текста в медицинском документе и получения семантических данных и информации [19].

Таким образом, преимущество этого типа графа заключается в том, что он фиксирует взаимосвязь между терминами. Однако недостатком является арифметическая сложность при сравнении графов, то есть он становится явно полиномиальным и имеет широкий диапазон параметров.

Существуют полуавтоматические концептуальные графические представления текста с использованием смеси существующих языковых ресурсов, таких как VerbNet и WordNet для выделения семантических частей [20].

Подводя итог, можно сказать, что богатая семантическая информация о материале может быть отражена в графе с помощью концептуального графа, однако остается фактом, что создание такого графа — непростая задача.

В таблице 1 приведен сравнительный анализ различных схем представления текста графом.

3. Области применения графового представления текста

3.1. Задача обнаружения плагиата

Для решения проблемы семантического обнаружения плагиата используются различные подходы, основанные на представлении графов. Один из методов, например, представляет каждое предложение внутри текстового документа в виде узла и объединяет все термины предложений в один узел. Заключенные узлы связаны друг с другом в соответствии с порядком предложений в текстовом документе. Извлеченные узлы затем объединяются в один большой узел на верхнем уровне, называемый узлом подписи темы. Сравнение между графами проводится на основе узлов подписи темы. И на основании сравнения делается вывод о плагиате [21].

3.2. Задача определения настроения текста

Методы представления текста с помощью графов можно использовать для анализа настроений текста [22]. Корпус текста известен как размеченный ориентированный граф, в котором слова являются узлами, а ребра указывают на синтаксические отношения между словами. При этом обход графа происходит с ограничением пути, в котором высокоуровневая информация о важных последовательностях направляет процесс обхода графа. Таким образом можно наблюдать улучшенную производительность и масштабируемость алгоритма обхода графа.

3.3. Применение графового представления текста в задачах машинного обучения

Графы полезны не только как организованные хранилища знаний. В современном машинном обучении они также играют ключевую роль. Например, можно использовать граф биологического взаимодействия для классификации роли белка [23], предсказать роль человека в сети совместной работы, предложить новых пользователей в социальной сети [24] или предсказать новое терапевтическое применение лекарственных молекул, структура которых может быть представлена в виде графа [25].

Для задач визуализации, кластеризации, классификации узлов и предсказания связей наиболее популярными являются включение узлов, и каждое из этих применений актуально для некоторых областей применения от вычислительных социальных наук до вычислительной биологии. В области обнаружения закономерностей и визуализации решается проблема просмотра графов в двумерном виде.

Классификация узлов может быть наиболее распространенным методом оценки встраивания узлов. Во многих случаях функция классификации представляет собой обучение с частичным привлечением учителя, в котором метки существуют только на небольшом количестве узлов для маркировки всего графа на основе этого небольшого начального набора. Популярные приложения обучения с частичным привлечением учителя для классификации узлов включают биологическую классификацию белков [26] и категории статей, изображений, веб-страниц или отдельных людей.

Для классификации текстов может использоваться такая структура, как графовые сверточные сети [27]. Также для представления гетерогенных концептуальных структур может использоваться модель нейронной сети с гетерогенным графом [28].

Заключение

Таким образом графовая модель является наиболее подходящим представлением текстового документа. В данной статье рассмотрены модели представления текста на основе графов, а также области их применения.

Графовая структура представляет собой узлы, обозначающие термины признаков, и ребра, обозначающие отношения между терминами. Отношения могут быть коокуррентными, грамматическими, семантическими или концептуальными. Отношения ребра для построения графа могут быть заменены взаимными отношениями между текстовыми сущностями. После того, как текстовый документ представлен в виде графа, к нему могут быть применены различные методы анализа графов.

Применение графового представления элементов текста обеспечивает обработку информации в различных областях, таких как кластеризация документов, классификация документов, разотождествление смысла слов, присоединение предложных фраз. Однако алгоритмы или методы на основе графов должны быть расширены, чтобы учесть требования и сложность приложений.

ЛИТЕРАТУРА

1. Jae-Yong Chang, Il-Min Kim Analysis and Evaluation of Current Graph-Based Text Mining Researches. Advanced Science and Technology Letters Vol. 42, 2013, pp. 100–103.
2. Овчинников В.А. Графы в задачах анализа и синтеза структур сложных систем.— М.: Изд-во МГТУ им. Н.Э. Баумана, 2014.— 423 с.
3. Мишунин О.Б., Савинов А.П., Фирстов Д.И. Проблемы, возникающие в интеллектуальных обучающих системах при оценке ответов на естественном языке // Современные проблемы науки и образования.— 2015.— № 2–2.
4. A.H. Osman, O.M. Barukub: Graph-Based Text Representation and Matching: A Review of the State of the Art and Future Challenges. IEEE Access: The Multidisciplinary Open Access Journal, Vol. 8, 2020, pp. 87562–87583.
5. Hassan S., Mihalcea R., Banea C., Random-Walk Term Weighting for Improved Text Classification. IEEE International Conference on Semantic Computing, ICSC-2007, 2007.
6. H. Balinsky, A. Balinsky, and S. Simske, Document Sentences as a Small World, in Proc. of IEEE SMC2011, pp. 9–12.
7. Bordag, S., Heyer, G., Quasthoff, U. Small worlds of concepts and other principles of semantic search. In T. Bhme, G. Heyer, H. Unger (Eds.), IICS, 2003, lecture notes in computer science Vol. 2877, pp. 10–19.
8. Cancho, R.F., Capocci, A., Caldarelli, G. Spectral methods cluster words of the same class in a syntactic dependency network. International Journal of Bifurcation and Chaos, 2007, 17(7), pp. 2453–2463.
9. J. Wu, Z. Xuan, and D. Pan Enhancing Text Representation for Classification Tasks with Semantic Graph Structures. International Journal if Innovative Computing, Information Control, Vol. 7, № 5(B), 2011, pp. 2689–2698.
10. Rada Mihalcea and Paul Tarau. TextRank: Bringing order into texts. Association for Computational Linguistics EMNLP-04, pp. 404–411.
11. Steyvers, M., Tenenbaum, J.B. The large-scale structure of semantic networks: Statistical analyses and a model of semantic growth. Cognitive Science, 2005, pp. 41–78.
12. Батура Т.В. Методы автоматической классификации текстов // Программные продукты и системы.— 2017. Т. 30. № 1. С. 85–99.
13. Cook, D.J. & Holder, L.B. Mining graph data, chapter Applications. John Wiley & Sons, 2006, pp. 345–468.
14. Deng, S., Sinha, A.P., & Zhao, H. Adapting sentiment lexicons to domain-specific social media texts. Decision Support Systems, 2016.
15. Freeman, L.C. Centrality in social networks conceptual clarification. Social Networks, Vol. 1, № 3, 1978, pp. 215–239.
16. Frery, J., Largeron, C., & Juganaru, M. Ujm at clef in author identification notebook for PAN at clef. CLEF 2014 Evaluation Labs and Workshop, Online Working Notes, Sheffield, UK, CEUR Workshop Proceedings, 2014, pp. 1042–1048.
17. Gelbukh, A. & Sidorov, G. Procesamiento automatico del espanol con enfoque en recursos lexicos grandes, chapter Tareas y aplicaciones de PLN. IPN, 2010, pp. 37–85.
18. Giannakopoulos, G., Karkaletsis, V., Vouros, G., & Stamatopoulos, P. Summarization system evaluation revisited: n-gram graphs. ACM Transactions on Speech and Language Processing, Vol. 5, № 3, 2008, pp. 1–38.
19. Dichiu, D. & Rancea, I. Using machine learning algorithms for author profiling in social media. CLEF 2016 Evaluation Labs and Workshop, Online Working Notes, Evora, Portugal, CEUR Workshop Proceedings, 2016, pp. 858–863.
20. Macke, S. & Hirshman, J. (2015). Deep sentence-level authorship attribution. Stanford University, pp. 1–7.
21. Harrington, P. Machine Learning in Action, chapter Support vector machines. Manning Publications Co., 2012, pp. 101–127.
22. Rosso, P., Rangel, F.M., Potthast, M., Stamatatos, E., Tschuggnall, M., & Stein, B. Overview of pan 2016 — new challenges for authorship analysis: Cross-genre profiling, clustering, diarization, and obfuscation. Proceedings of CLEF PAN2016, Lecture Notes in Computer Science, 2016, pp. 332–350.
23. Samatova, N.F., Hendrix, W., Jenkins, J., Padmanabhan, K., & Chakraborty, A. Practical graph mining with R, chapter Introduction. Chapman & Hall/CRC, 2013, pp. 1–7.
24. Samatova, N.F., Hendrix, W., Jenkins, J., Padmanabhan, K., & Chakraborty, A. Practical graph mining with R, chapter Frequent subgraph mining. Chapman & Hall/CRC, 2013, pp. 180–186.
25. Samatova, N.F., Hendrix, W., Jenkins, J., Padmanabhan, K., & Chakraborty, A. Practical graph mining with R, chapter Link analysis. Chapman & Hall/CRC, 2013, pp. 110–121.
26. Sharma, A. & Shubhamoy, D. An artificial neural network-based approach for sentiment analysis of opinionated text. Proceedings of the ACM Research in Applied Computation Symposium, New York, USA, ACM, 2012, pp. 37–42.
27. Stanczyk, U. & Cyran, K.A. Machine learning approach to authorship attribution of literary texts. International journal of applied mathematics and informatics, Vol. 1, № 4, 2007, pp. 151–158.
28. Stojanovski, D., Strezoski, G., Madjarov, G., & Dimitrovski, I. Finki at semeval-2016 task 4: Deep learning architecture for twitter sentiment analysis. Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016), ACL, 2016, pp. 149–154.

Акбашева Галина Амировна (galina_akbash@mail.ru), Тлупов Ислам Заурбекович (tlupovislam@gmail.com).

Журнал «Современная наука: актуальные проблемы теории и практики»