Модель поиска схожих документов в семантической библиотеке

Атаева О.М., Серебряков В.А.,Тучкова Н.П.

Вычислительный центр им. А.А. Дородницына ФИЦ ИУ РАН

Аннотация. Рассматривается задача нахождения наиболее релевантных документов в результате расширенного и уточненного запроса. Для этого предлагается модель поиска и механизм предварительной обработки текста, а также совместное использование поисковой системы и нейросетевой модели, построенной на основе индекса с помощью алгоритмов word2vec для генерации расширенного запроса с синонимами и уточнения результатов поиска на основе подбора похожих документов в цифровой семантической библиотеке. В работе исследуется построение векторного представления документов на основе абзацев применительно к массиву данных цифровой семантической библиотеки LibMeta. Решалась задача обогащения пользовательских запросов синонимами. При построении модели поиска совместно с алгоритмами word2vec использовался подход «сначала индексация, затем обучение», чтобы охватить большее количество информации и выдать более точные результаты.

Ключевые слова: модель поиска, word2vec, синонимы, запрос, расширение запроса

Search model for similar documents in the semantic library

O.M. Ataeva[*][0000-0003-0367-5575], **V.A. Serebryakov**[**] [0000-0003-1423-621X], **N.P. Tuchkova**[***][0000-0001-6518-5817]

Dorodnicyn Computing Center FRC CSC of RAS
[*]oli@ultimeta.ru, [**]serebr@ultimeta.ru, [***]natalia_tuchkova@mail.ru

Abstract. The problem of finding the most relevant documents as a result of an extended and refined query is considered. For this, a search model and a text preprocessing mechanism are proposed, as well as the joint use of a search engine and a neural network model built on the basis of an index using word2vec algorithms to generate an extended query with synonyms and refine search results based on a selection of similar documents in a digital semantic library. The paper investigates the construction of a vector representation of documents based on paragraphs in relation to the data array of the digital semantic library LibMeta. Each piece of text is labeled. Both the whole document and its separate parts can be marked. The problem of enriching user queries with synonyms was solved, then when building a search model together with word2vec algorithms, an approach of "indexing first, then training" was used to cover more information and give more accurate results.

Keywords: search model, word2vec, synonyms, query, query extension

1. Введение

Цели формирования научного информационного пространства включают сохранение и извлечение знаний [1]. Благодаря глобальной идее накопления знаний в цифровом виде, появилась возможность получать наукометрические данные, рейтинги ученых и изданий. В какой-то момент обладание базой знаний стало выгодно с научной и коммерческой точки зрения. Поиск информации практически однозначно воспринимается сегодня как запрос в информационной системе, что нередко приводит к извлечению недостоверной информации [2]. Проблемы появления мусорной, фейковой, неверной информации в сети можно определить как следствие ошибочных семантических связей в базах данных. Это свидетельствует о том, что на предварительной обработке документы были неправильно размечены. Для научных работ, помещенных с ошибочными индексами в базу данных, это означает, что такие работы могут быть не найдены специалистами в некоторых предметных областях и не процитированы. В этом контексте особую роль играет предварительная обработка данных, которые, обретя определенную структуру, в дальнейшем будут извлекаться в качестве знаний [3].

Ставится задача связать модель поиска с причинно-следственным механизмом генерации данных и таким образом отследить семантические связи. То есть определяется предметная область, соответствующий ей словарь со связями, на основе которого связи между терминами определяют связи между документами, статьями, авторами и т.д. Для реализации такой схемы необходима предварительная обработка текстов и применение алгоритмов сравнения, оценки схожести. Современные подходы к решению задачи поиска достоверной научной информации предполагают применение методов машинного обучения для ранжирования результатов поиска, нейронных сетей для обработки текста и других методов интеллектуального анализа данных [4, 5].

Настоящая работа посвящена вариантам разрешения проблемы предоставления релевантных результатов в отношении информационной потребности пользователя. Для этого предлагается модель поиска и механизм предварительной обработки текста, а также совместное использование поисковой системы и нейросетевой модели, построенной на основе индекса с помощью алгоритмов word2vec для генерации расширенного запроса с синонимами и уточнения результатов поиска на основе подбора похожих документов в цифровой семантической библиотеке LibMeta [6].

2. Модель поиска

Построение модели поиска основывается на трех основных ключевых пунктах, а именно:

- Представление документа в формате, пригодном для поиска. В нашем случае для подготовки документов проводилась предобработка полных текстов для удаления издательской разметки и выделения основных частей текста. Затем был создан полнотекстовый индекс документов, который позволяет эффективно загружать и хранить данные и обеспечивает быстрый доступ к ним.
- Представление запросов в формате, позволяющем выражать информационные потребности пользователя. В качестве запросов используются запросы, написанные на естественном языке, которые системой могут быть обогащены синонимами.
- Оценка соответствия документа запросу является субъективной и зависит от применяемого метода.

Основная проблема любой поисковой модели – предоставление релевантных результатов в отношении информационной потребности пользователя: от анализа запроса до ранжирования результатов поиска. Одним из современных подходов разрешения этой проблемы является использование нейронных сетей для обработки текста. Текст представляет

собой пример данных, которые можно разобрать на более мелкие структуры такие, как параграфы, предложения, слова и т.д. в зависимости от специфики. Такой подход к обработке текста позволяет захватывать семантику текста, так как тесно связанные слова или фрагменты текста встречаются в одинаковом контексте и лежат рядом в векторном пространстве. Используемая в работе модель поиска опирается на векторное представление слов и документов построенное с помощью алгоритма нейронной сети word2vec [7,8,9].

Интеграция нейронной сети и индекса может выполняться следующими способами:

- сначала обучение на корпусе текстов, затем индексация текстов и совместное использование обученной модели и индекса при поиске;
- сначала индексация, затем обучение на индексированных данных и совместное использование при поиске;
- сначала обучение, затем извлечение/создание полезных ресурсов обученной сетью, и потом индексация всех ресурсов и новых и исходных.

В работе решалась задача обогащения пользовательских запросов синонимами. Для этого в библиотеке LibMeta использовался подход «сначала индексация, затем обучение». При этом с одной стороны ставилась задача предоставить больше более точных результатов на основе расширенных запросов [10,11,12]. С другой - при использовании расширенной версии word2vec совместно с поисковой системой LibMeta появляется возможность давать пользователям более «умные рекомендации» на основе найденных документов. Такой подход к совместному использованию индекса поисковой системы и нейросети позволяет получать релевантные модели и функции ранжирования, которые хорошо адаптируются к базовым данным. Версию модели, построенную на поисковом индексе LibMeta с помощью алгоритмов word2vec, далее будем сокращенно называть wsgMath [13].

Совместное использование полнотекстового индекса и модели делают возможным расширение исходного запроса синонимами. Расширение запросов синонимами без wsgMath требует предварительно составленных словарей синонимов.

Замечание 1. Можно использовать такие ресурсы, как WordNet или RuWordNet, но основная проблема в том, что синонимы из предварительно составленных словарей не привязаны к индексируемым данным и их использование, как было проверено, не улучшает результаты.

На рис. 1 представлены основные шаги формирования модели для генерации синонимов запроса в LibMeta. Строка запроса, поступающая из интерфейса полнотекстового поиска, проходит через блок *Анализатора*. В *Анализаторе* строка разбивается на слова, проводится их анализ и

преобразование. Из модели wsgMath извлекаются и фильтруются синонимы к словам, что позволяет сформировать расширенный запрос, с помощью которого из полнотекстового индекса извлекаются соответствующие документы.

Применение расширенной версии word2vec (doc2vec или paragraph2vec, в разных источниках по - разному) позволяет ввести дополнительный элемент, такой, как *метка фрагмента текста* или всего документа и, основываясь на векторах этих меток, подбирать похожие документы не только по точному совпадению ключевых слов или терминов, но *основываясь на контексте* отдельных фрагментов или всего документа.

Замечание 2. *Метка фрагмента текста* используется для выдачи близких по смыслу документов, которые не попадают в поисковую выдачу, но могут представлять интерес для пользователя.

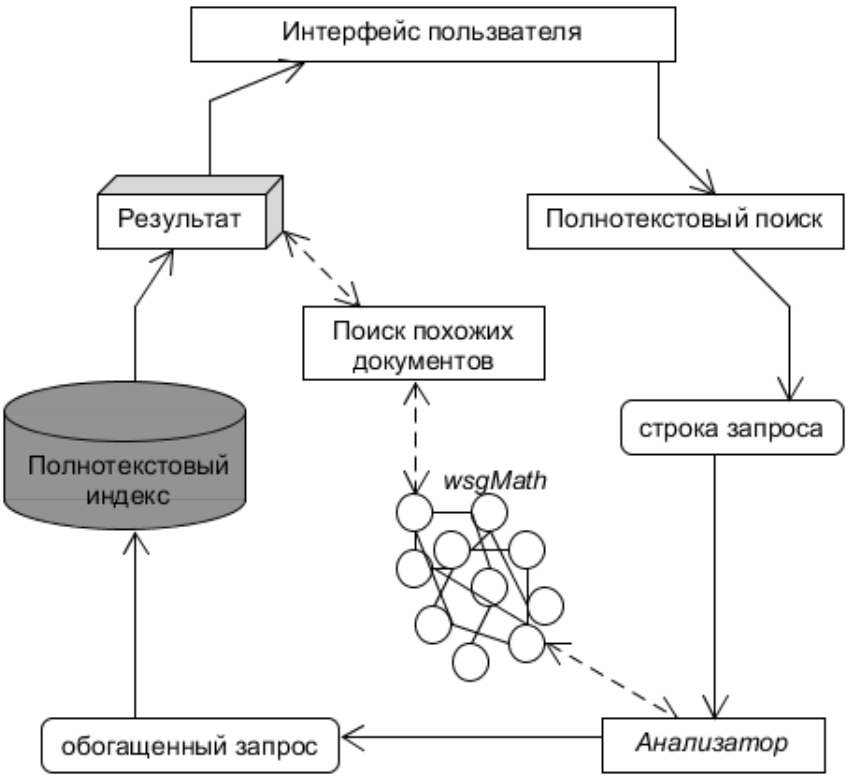

Рис. 1. Совместное использование поисковой системы и нейросетевой модели, построенной на основе индекса с помощью алгоритмов word2vec для генерации расширенного запроса с синонимами и уточнения результатов поиска на основе подбора похожих документов.

3. Построение и исследование модели векторного пространства поисковой системы

3.1. Предобработка статей

Одним из необходимых этапов подготовки данных к их загрузке в определенных текстовых форматах в уже подготовленную инфраструктуру данных является *препроцессинг* и *очистка* этих данных. В нашем случае данные предоставлялись в файлах в формате tex.

В связи с тем, что файлы были оформлены в разных стилях и команды обозначались разными именами, необходимо было заменить все авторские тэги на стандартные, очистить документы от специальных символов и неизвестных тэгов. При этом совсем избежать ручной обработки не удалось, но по крайней мере, ее удалось свести к минимуму.

Модуль предварительной обработки выполнен на языке программирования Python вместе с интеграцией open-source библиотеки TexSoup версии 2015 года и разбит на следующие блоки:

- очистка документа,
- преобразование статьи в древовидное представление,
- обработка всех узлов дерева, запись исправленного документа.

На рис. 2 представлены основные этапы предварительной обработки текстов.

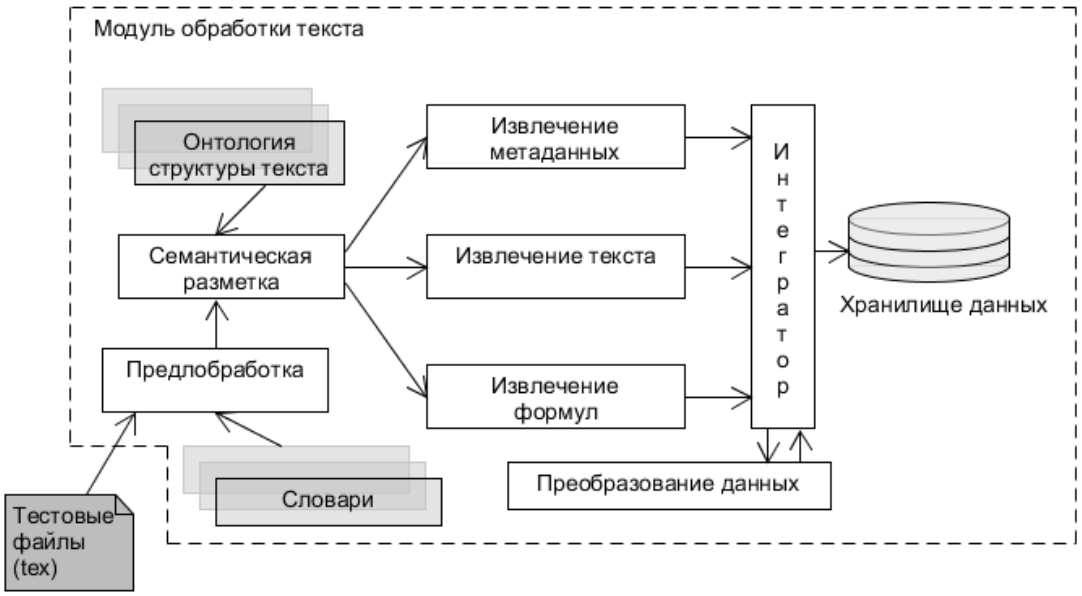

Рис. 2. Схема предварительной обработки текстов

3.2. Построение индекса и обучение word2vec

Далее на основе предобработанных статей был построен *индекс полных текстов* на базе библиотеки поиска с открытым исходным кодом Apache Lucene, написанным на Java. Этот индекс используется подсистемой полнотекстового поиска библиотеки и он же использовался для обучения алгоритма и извлечения контекстов. Ниже приведены примеры контекстов, которые встречаются для термина «задача коши», извлекаемых из полнотекстового поискового индекса.

…разрешимость ***задача коши*** *уравнение аллер пространство непрерывный ограниченный функция сводить разрешимость абстрактный* ***задача коши*** *банаховый пространство непрерывный ограниченный функция ось…*

…исследовать фундаментальный решение обобщенный ***задача коши*** *один гиперболический оператор изученный свойство фундаментальный решение…*

В Таблицах 1-2 приводятся примеры семантического анализа контента семантической библиотеки на основе модели wsgMath. В Таблице 1 приведены примеры слов со связями между словами (в первой строке стоит главное слово, в столбцах ниже – выявленные).

Таблица 1. Все связи слова

Коши	***задача***	***теорема***	***уравнение***
Риман	проблема	Лемма	задача
краевой	система	предложение	дифференциальный
постановка	уравнение	ограниченный	тождество
решать	вопрос	слоить	система
	решение		

В Таблице 2 приведены примеры синонимов, определенных по частям речи и коэффициенту близости слов (в нашем случае отбирались существительные при $\delta = 0,45$).

Таблица 2. Связи слова с учетом частей речи

коши	***задача***	***теорема***	***уравнение***
	уравнение	лемма	задача
	система	предложение	тождество
	решение		

4. Примеры (запрос с синонимами)

4.1. Расширение синонимами

Рассмотрим термин «*задача коши*», который тоже состоит из двух слов, «*задача» и «коши*», *каждое из которых имеет собственные синонимы, которые представлены в* Таблице 3. В третьем столбце представлен контекст термина как одной единицы. Извлекая его *синонимы*, видно, что список состоит из слов, куда попадает прилагательное *краевой*, которое также встречается в *синонимах отдельных слов в первых двух столбцах.*

При этом сам термин «*задача коши»* имеет следующие *синонимы: «уравнение коши», «система коши», «решение коши», которые были определены на основе высоких оценок близости следующих пар синонимов:*
sim (задача, решение) = 0.87,
sim (задача, уравнение) = 0.84,
sim (задача, система) = 0.71.

Замечание 3. При построении синонимичных терминов не использовались синонимы слова *коши*, так как оно было определено как именованная сущность (*Коши*) на основе словаря, который включает в себя список персон, упоминания о которых встречаются в математической энциклопедии. Но при этом отметим, что в *синонимы коши* попало *риман (Риман)*.

На основе этих синонимов были сформированы следующие предложения, которые получены в соответствии с рассматриваемым паттерном на модели wsgMath *[краевая задача коши, краевое уравнение коши, краевая система коши, краевая решение коши]*.

Таблица 3. *CS-synonyms* для термина «задача коши» и его слов

задача	***коши***	***задача коши***
уравнение	риман	постоянный
система	краевой	определять
решение		краевой
краевой		

Начало списка результатов поиска выглядит как

1. О положительном радиально-симметрическом решении задачи Дирихле для одного нелинейного уравнения и численном методе его получения

score = 0.9809638

2. О корректности краевой задачи на прямой для трех аналитических функций

score = 0.9587569

3. О положительном радиально-симметрическом решении задачи дирихле для одного нелинейного уравнения и численном методе его получения

score = 0.9512307

При этом искомый документ располагается на второй позиции

4.1. Поиск похожих документов

Рассмотрим пример использования элемента *Метка фрагмента текста*. Посмотрим на процесс *ранжирования* документов на основе модели wsgMath при поиске схожих документов. Когда документ поступает в систему, то извлекается его текущее векторное представление, выполняется поиск и возвращаются *метки ближайших документов, косинусное расстояние которых превышает некоторый порог, определенный экспериментально как 0,6.* Ниже представлен результат работы на примере документа, который в предыдущем примере являлся искомым. В качестве ближайших к нему было найдено 9 документов, косинусное расстояние у которых превышало 0.6.

1. Некоторые классы сингулярных интегральных уравнений разрешаемые в замкнутой форме

cosineSimilarity=0.8136491179466248

2. Краевая задача Римана для полуплоскости с коэффициентом экспоненциально убывающим на бесконечности

cosineSimilarity=0.8028532266616821

3. Алгоритм построения квазирегулярного асимптотического представления решения сингулярно возмущенных линейных многоточечных краевых задач с быстрыми и медленными переменными

cosineSimilarity=0.7246567010879517

4.2. Измерение качества похожих документов на основе кодов классификации документов

Один из вариантов оценки качества похожих документов - это использование классификаторов.

Примеры для случаев:

1. Документы, поступающие в систему, размечены кодами классификаторов MSC и УДК. В этом случае при выявлении документов, косинусное расстояние у которых превышало

заданный порог, можно сравнивать коды классификаторов и устанавливать соответствие MSC и УДК. Если коды УДК отличаются у схожих документов, то можно их указать как смежные предметные области (приложения результатов, междисциплинарные исследования и пр.).

2. Документы не снабжены кодами, но ключевые слова соответствуют предметной области и в словаре (тезаурусе, энциклопедии) есть коды классификаторов. В этом случае сравниваются коды ключевых слов и документам приписываются соответствующие коды.

Заключение и дальнейшие исследования

Показано, что предварительная обработка входных массивов данных (текстов научных статей) позволяет учесть в дальнейшем дополнительные семантические связи документов. В результате использования механизма интеграции нейронной сети и индекса получены варианты поисковой модели для получения релевантных документов при поисковом запросе с заданной точностью. Предложенная модель поиска позволяет также устанавливать соответствие кодов классификаторов для близких документов, находить синонимы при контекстном сравнении и ранжировать документы на основе метки фрагмента.

Работа представлена в рамках выполнения темы госзадания «Математические методы анализа данных и прогнозирования» ФИЦ ИУ РАН и при частичной поддержке Российского фонда фундаментальных исследований (проект №20-07-00324).

Литература

1. Biswas, G., Bezdek, J., and Oakman, R. L.: A knowledge-based approach to online document retrieval system design. In Proc. ACM SIGART Int. Symp. Methodol. pp. 112–120. Intell. Syst. (1986).
2. Furnas, G. W., Landauer, T. K., Gomez, L. M., and Dumais, S. T.: The vocabulary problem in human-system communication. Commun. ACM, 30(11), 964–971 (1987).
3. Гаврилова Т.А., Хорошевский В.Ф. Базы знаний интеллектуальных систем. СПб.: Питер, 2000. 384 с.
4. Профессиональный информационно-аналитический ресурс, посвященный машинному обучению, распознаванию образов и интеллектуальному анализу данных. http://www.machinelearning.ru/.
5. Мак-Каллок У. С., Питтс В. Логическое исчисление идей, относящихся к нервной активности // Автоматы / Под ред. К. Э. Шеннона и Дж. Маккарти. — М.: Изд-во иностр. лит., 1956. — С. 363—384. (Перевод английской статьи 1943 г.

6. Атаева О.М., Серебряков В.А, Онтология цифровой семантической библиотеки LibMeta //Информатика и её применения. 2018. Т. 12. №. 1. С. 2-10.
7. Mikolov T., Chen K., Corrado G., Dean J. Efficient Estimation of Word Representations in Vector Space // Proceedings of Workshop at ICLR, 2013.
8. Mikolov T., Yih W.T., Zweig C. Linguistic Regularities in Continuous Space Word Representations // Proceedings of NAACL HLT, 2013.
9. Le Q., Mikolov T. Distributed Representations of Sentences and Document // International Conference on Machine Learning, 2014. pp. 1188-1196
10. Voorhees, E. M.: Query expansion using lexical-semantic relations. 17th Annu. Int. ACM SIGIR Conf. Res. Develop. Inf. Retr., Dublin, Ireland (1994).
11. Buckley C., Salton G., Allan J., and Singhal A. Automatic query expansion using SMART: TREC 3, presented at the 3rd Text Retr. Conf. (TREC), 1995.
12. Efthimiadis E. N. Query expansion // Annu. Rev. Inf. Sci. Technol., V. 31, no. 5, pp. 121-187, 1996.
13. Ataeva O.M., Sererbryakov V.A., Tuchkova N.P. Using Applied Ontology to Saturate Semantic Relations // Lobachevskij Journal of Mathematics. 2021. Vol. 42. N. 8. pp. 1776–1785.

References

1. Biswas, G., Bezdek, J., and Oakman, R. L.: A knowledge-based approach to online document retrieval system design. In Proc. ACM SIGART Int. Symp. Methodol. pp. 112–120. Intell. Syst. (1986).
2. Furnas, G. W., Landauer, T. K., Gomez, L. M., and Dumais, S. T.: The vocabulary problem in human-system communication. Commun. ACM, 30(11), 964–971 (1987).
3. Gavrilova T A Horoshevskij V F Bazy znanij intellektualnyh sistem SPb Piter 2000 384 c
4. Professionalnyj informacionno analiticheskij resurs posvyashchennyj mashin nomu obucheniyu raspoznavaniyu obrazov i intellektualnomu analizu dannyh http www machinelearning ru
5. Mak Kallok U S Pitts V Logicheskoe ischislenie idej otnosyashchihsya k nervnoj aktivnosti Avtomaty Pod red K EH SHennona i Dzh Makkarti M Izd vo inostr lit 1956 S 363 384 Perevod anglijskoj stati 1943 g.
6. Атаева О.М., Серебряков В.А, Онтология цифровой семантической библиотеки LibMeta //Информатика и её применения. 2018. Т. 12. №. 1. С. 2-10.
7. Mikolov T., Chen K., Corrado G., Dean J. Efficient Estimation of Word Representations in Vector Space // Proceedings of Workshop at ICLR, 2013.
8. Mikolov T., Yih W.T., Zweig C. Linguistic Regularities in Continuous Space Word Representations // Proceedings of NAACL HLT, 2013.

9. Le Q., Mikolov T. Distributed Representations of Sentences and Document // International Conference on Machine Learning, 2014. pp. 1188-1196
10. Voorhees, E. M.: Query expansion using lexical-semantic relations. 17th Annu. Int. ACM SIGIR Conf. Res. Develop. Inf. Retr., Dublin, Ireland (1994).
11. Buckley C., Salton G., Allan J., and Singhal A. Automatic query expansion using SMART: TREC 3, presented at the 3rd Text Retr. Conf. (TREC), 1995.
12. Efthimiadis E. N. Query expansion // Annu. Rev. Inf. Sci. Technol., V. 31, no. 5, pp. 121-187, 1996.
13. Ataeva O.M., Serebryakov V.A., Tuchkova N.P. Using Applied Ontology to Saturate Semantic Relations // Lobachevskij Journal of Mathematics. 2021. Vol. 42. N. 8. pp. 1776–1785.