Научная статья
УДК 81’32+004.8

Система для интеграции знаний в текстовом формате

А. А. Харламов
Институт высшей нервной деятельности и нейрофизиологии Российской академии наук. Москва, Россия
kharlamov@ihna.ru

Аннотация. В статье представлены биологические предпосылки, теоретические соображения, алгоритмы и краткое описание прототипа системы для интеграции знаний в текстовом формате. Цель – создание среды для хранения общественного знания, не подверженной необходимости перемасштабирования в связи с увеличением объема хранимой информации, изменения ее структуры вследствие динамики развития знаний – исчезновения старых и появления новых разделов. Эта среда является гипертекстовым представлением и включает, помимо исходных текстов также семантическую сеть, характеризующую содержание этих текстов.

Ключевые слова: среда для представления знаний, представление текстовой информации, автоматическая обработка текста, семантическая сеть как представление содержания текста, программа TextAnalyst

Для цитирования: Харламов А. А. Система для интеграции знаний в текстовом формате // Вестник Московского государственного лингвистического университета. Гуманитарные науки. 2014. Вып. 10 (891). С. 119–125.

Original article

System for Integrating Knowledge in Text Format

Alexander A. Kharlamov
Institute of Higher Nervous Activity and Neurophysiology of the Russian Academy of Sciences, Moscow, Russia
kharlamov@ihna.ru

Abstract. The paper presents biological background, theoretical considerations, algorithms and a brief description of a prototype system for integrating knowledge in text format. The purpose of the system is to create an environment for storing public knowledge that does not require rescaling caused by an increase in the volume of stored information and changes in its structure due to the dynamics of knowledge development, that is, the disappearance of old sections and the appearance of new ones. The system is represented as a hypertext and includes, in addition to the source texts, also a semantic network that characterizes the content of these texts.

Keywords: environment for knowledge representation, representation of text information, automatic text processing, semantic network as a representation of text content, TextAnalyst program

For citations: Kharlamov, A. A. (2024). System for integrating knowledge in text format. Vestnik of Moscow State Linguistic University. Humanities, 10(891), 119–125. (In Russ.)

ВВЕДЕНИЕ

Современность характеризуется экспоненциальным ростом объема продуцируемых человечеством знаний, доступность к которым у рядового пользователя обратно пропорциональна их объемам. Назрела необходимость централизации доступа к знаниям, которая сопровождается необходимостью интеграции самих знаний. По крайней мере, при их распределенном хранении за счет такой организации представления этих знаний, которая позволяет пользователю эффективно их получать.

Предпринимаются попытки реализовать подобные принципы в рамках отдельных предметных областей [Информационная система анализа научной деятельности (ИСАНД) в области теории управления, 2024], что возможно при использовании современных подходов к структурированию хранимой информации, которое должно реализовываться автоматически.

Имеющиеся предложения на эту тему, касающиеся отдельных предметных областей, не гарантируют масштабирования как при наращивании числа охватываемых предметных областей, так и при увеличении объема хранимых знаний в рамках отдельных предметных областей в силу громоздкости механизма хранения. В настоящий момент в этот механизм заранее закладывается структура предметной области; изменение предметной области предполагает полную переработку этой структуры под другие знания. Поэтому необходим иной подход, более естественный относительно хранимых знаний, как с точки зрения их содержания (универсальный с точки зрения содержания), так и с точки зрения их объема. Кстати, в упомянутых работах пока что речь идет об автоматической обработке и хранении лишь текстовой информации, да и то автоматическая обработка не распространяется на содержание текстов, а только на выходные данные (название публикации, аннотация).

В настоящий момент системы, которые интеллектуально решают возложенные на них задачи, работают успешно, более или менее точно моделируя соответствующие функции человеческого мышления: человек остается единственной информационной системой, эффективно решающей все возложенные на него интеллектуальные задачи. Поэтому при построении эффективно работающей системы получения, обработки и хранения информации, естественно посмотреть, как человек решает задачи сбора, структуризации, хранения, выдачи (on-demand) разномодальной информации. Надо заметить, что эта эффективность использования вытекает из интегральности представляемой информации: наряду с текстовой модальностью в системе должна быть представлена информация других модальностей, в первую очередь – зрительной.

СИСТЕМА ДЛЯ ИНТЕГРАЦИИ ЗНАНИЙ

Поговорим о системах хранения текстовой информации. Система для интеграции знаний – цифровой аналог обычной библиотеки. Таких сейчас, наверное, очень много. Здесь не предпринимается попытка теоретического вычисления качества существующих систем, но попытка взглянуть на предмет с точки зрения возможности создания идеальной, по мнению пользователя, цифровой библиотеки.

Желаемое более или менее понятно. Это структурированное цифровое текстовое хранилище, доступ к текстам которого осуществляется по ассоциации [Харламов, 2017]. Возможно, можно попытаться описать процедуру получения нужных единиц хранения на примере, точнее, на двух примерах. (1) Первый класс задач – поиск текстов в хранилище на заданную тему. (2) Второй класс задач – поиск встречных текстов (как в хранилище, так и вне его).

Первый класс задач более или менее понятен: предметная область задается примером (коротким текстом). В хранилище она уже структурирована, или структурируется на проходе. В отличие от представленного в статье «Информационная система анализа научной деятельности (ИСАНД) в области теории управления» (2024) структурирование осуществляется по мере заполнения хранилища [там же]. Структура выявляется для конкретного корпуса текстов, как это делается в мозге (описано ниже).

Под *встречным текстом* понимается текст, который соединяет два разрозненных корпуса текстов таким образом, что при наличии этого встречного текста два корпуса становятся частями единого третьего объединенного корпуса [Леонтьева, 2006].

Поиск текста

Задача поиска текста в цифровом хранилище сравнительно (теоретически) проста. Правда, в отличие от имеющихся поисковых систем в нашем случае в качестве запроса на поиск мы нуждаемся в тексте, который описывает контекст поиска. В этом случае исходный текст как текст запроса собирает из хранилища все тексты, совокупность которых содержит текст запроса в качестве некоторой ядерной части.

Далее это множество текстов структурируется, выбирается некоторая бахрома текстов, содержащая заданный (параметр задается заранее) объем информации, содержащая и ядерную часть и отличающаяся по объему в большую сторону на заданное число процентов. В структуре этой части корпуса текстов выявляются концепты, которые

и представляют интерес (которые в текстах связаны с заданным во входном текстовом запросе перечнем концептов). И происходит анализ этих новых концептов и их связей как между собой, так и с исходным, и с дополнительным множествами концептов.

Поиск встречного текста

Поиск встречного текста – более сложная задача. Имеется два корпуса текстов, к которым ищется встречный текст. Задача предварительно выглядит следующим образом. Поскольку объединение двух корпусов встречным текстом осуществляется по каким-то концептам семантической сети, семантическая сеть встречного текста должна объединить семантические сети корпусов так, что часть концептов встречного текста совпадает с концептами первого из двух объединяемых корпусов текстов, а другая часть концептов встречного текста – совпадает с концептами второго корпуса текстов.

Теоретически задача решается следующим образом. Анализируются оба сшиваемых корпуса текстов. Выявляются принадлежащие им концепты. Далее (подбором) выявляются тексты из внутреннего или внешнего корпусов, имеющие в составе те и другие концепты (или какую-то их часть), обрабатываются вместе оба сшиваемых корпуса в совокупности с найденным встречным текстом. И экспертно подбирается лучший вариант с различными подходящими встречными текстами.

И в том, и в другом случае возможно реализовать систему как вопросно-ответную, в которой и искомый текст (как в случае 1), и встречный текст (как в случае 2) ищутся по запросу пользователя. А запросом может быть и текст из одного предложения, и более чем из одного предложения также.

ОБРАБОТКА И ХРАНЕНИЕ ИНФОРМАЦИИ В МОЗГЕ ЧЕЛОВЕКА

Поэтому, поговорим, во-первых, о представлении информации разных модальностей, во-вторых – о механизмах ее использования, как это реализуется в мозге человека. Понимая то и другое, мы сможем дать рекомендации по реализации системы для эффективного хранения знаний в текстовом формате. А понимая это, можно попытаться прогнозировать пути ее развития с учетом развития человечества в этом направлении.

В мозге человека модель мира представлена в двух ипостасях: в виде иерархии словарей уровнеобразующих единиц (как языка – для языковой модели, так и квазиязыков – для экстралингвистической модели), с одной стороны, а, с другой – в виде множества шаблонов ситуаций. Иерархии словарей представлены в колонках коры, а шаблоны ситуаций – в ламелях гиппокампа. Для описания целевой системы достаточно иметь иерархию словарей, представленных в колонках коры полушарий. Задача хранения текстовой информации сопряжена лишь с языковой иерархией.

В случае работы с текстовой информацией (языковая модель) речь идет о словарях уровнеобразующих элементов языка – от словаря графем на нижнем уровне (если текст письменный, или печатный) до словаря допустимой попарной сочетаемости корневых основ слов – на верхнем семантическом уровне. Эти пары корневых основ объединяются виртуально в однородную семантическую сеть, весовые характеристики элементов и связей которой указывают на их ранги (значимость) в конкретных текстах.

Такое же сетевое представление формируется и в экстралингвистической части модели мира (например, для представления зрительной информации), и эти два представления, описывающие лингвистически и экстралингвистически один и тот же мир, объединяются в единое семантическое представление. Тогда любой текст (или квазитекст) может быть представлен такой его семантической сетью (в случае анализа текста языка формируется только языковая сеть). Эти сети можно сравнивать друг с другом, выявляя степень их подобия, и таким образом, классифицировать. Эти сети можно кластеризовать, т. е. корпус текстов может быть разбит на подкорпуса в соответствии с их внутренней структурой. Из этих сетей можно выделять минимальный древовидный подграф, который является оглавлением текста. На основе такой сети текста выявляется некое смысловое ядро текста, которое можно считать его аннотацией [Kharlamov, Pilgun, 2020].

СИСТЕМА ДЛЯ ИНТЕГРАЦИИ ЗНАНИЙ В ТЕКСТОВОМ ФОРМАТЕ

Попробуем представить эффективную систему для обработки и хранения текстовой информации, которая при необходимости могла бы быть дополнена подсистемой для обработки и хранения экстралингвистической информации, с одинаково удобной визуализацией хранящейся в системе информации как на уровне оглавления, так и на уровне представления отдельных статей.

Естественно, ничего другого на первый взгляд (помимо отдельных технических решений) кроме как повторения архитектуры мозга в части получения, обработки, хранения и выдачи специфической информации придумать не удается. Поэтому рассмотрим еще раз архитектуру системы

обработки специфической информации, реализованной в мозге человека как субстрате. В рамках этой архитектуры только модуль визуализации содержания существенно отличается от естественного прототипа: вместо написания текста как процесса управления эффекторикой кисти доминантной руки реализуется вывод цифрового текста на экран. В состав системы также входит модуль хранения полных текстов статей.

Функциональность системы

Система для интеграции знаний должна реализовать следующую функциональность: (1) анализ текстов с формированием семантической сети как отражающей содержание текста, так и являющейся основой для реализации функциональности системы; (2) хранение как корпуса текстов, так и семантических сетей текстов корпуса; (3) формирование аннотаций текстов для удобного знакомства пользователей с базой текстов; (4) кластеризация корпуса текстов на предметные области; (5) сравнение текстов по смыслу; (6) классификация текстов – отнесение к предметным областям; (7) поиск текста (множества текстов) по запросу; (8) поиск встречного текста для двух и более текстов.

Архитектура

Архитектура системы получения, анализа, хранения, выдачи текстовой информации включает в свой состав так называемый лингвистический процессор, где подробности преобразования звучащего текста в рукописный или печатный и в обратную сторону выносятся за рамки этой статьи. Рассматривается уже оцифрованный текст без исследования механизма оцифровки.

Архитектура системы включает в себя много лингвистических подробностей разных языковых уровней; не все из них необходимо использовать (или, по крайней мере, описывать) при реализации системы. В основе представления знаний в системе рассматривается семантическое представление текста как однородной взвешенной семантической сети, множество вершин которой соотносятся с концептами языка с их весовыми характеристиками, представленными в тексте (извлекаемыми из текста), а дуги соответствуют ассоциативным связям этих концептов друг с другом в рамках этого текста, также взвешенным их весовыми характеристиками [Харламов, 2017] (см. рис. 1).

Программная система для интеграции знаний в текстовом формате содержит блок первичной обработки (1), лингвистический и семантический процессоры. Лингвистический процессор (2) состоит из словарей: (4) слов-разделителей, (5) служебных слов, (6) общеупотребимых слов, а также (7) флективных и (8) корневых морфем. Семантический процессор (3), в свою очередь, содержит: (9) блок отсылок в текст, (10) блок формирования семантической сети, (11) блок хранения семантической сети, (12) блок выделения понятий и (13) блок управления.

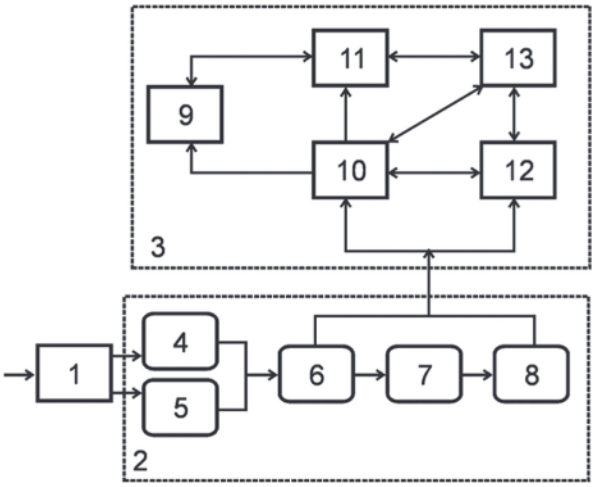

Рис. 1. Программная система для интеграции знаний в текстовом формате

Для формирования подобной сети требуется информация семантического и других языковых уровней: морфемного и лексического. Опишем более подробно обработку информации на этих трех уровнях, результатом которой является направленная однородная взвешенная семантическая сеть, представляющая содержание текста.

Для удобной работы с текстом его необходимо очистить от нетекстовой информации, а также удалить из него текстовую информацию, не участвующую в дальнейшей обработке. В дальнейшей обработке не участвуют служебные слова (артикли, предлоги, местоимения, если они не раскрываются в процессе раскрытия анафорических ссылок). В случае использования однородной семантической сети значение глагола не работает, так как связи в сети – однородные, поэтому глаголы также выбрасываются из рассмотрения как рабочие слова. Их наличие заменяется просто на связь.

Формально для формирования семантической сети морфемный уровень не нужен. Все словоформы участвуют в формировании семантической сети на равных правах. Это станет понятно из дальнейшего изложения. Но для получения более представительной (более робастной) семантической сети – сети, более удобной для реального анализа текстов – полезно контрастировать лексическое содержание текста:

привести словоформы к их корневым основам. При этом формирующаяся сеть становится более грубой, но зато более наглядной.

Итак, первый шаг анализа текста – стемминг. Все слова текста сводятся к их корневым основам.

На лексическом уровне формируется перечень корневых основ с частотой их употребления в тексте.

На семантическом уровне представлена попарная допустимая сочетаемость корневых основ слов, как они перечислены в предложениях текста. Результатом первичного семантического анализа текста является перечень пар корневых основ, как они встречаются в предложениях текста, с их весовыми характеристиками (частота встречаемости).

Ранг вершины (концепта) в тексте определяется не его частотой встречаемости в тексте, а объемом той части семантической сети, которая удерживается концептом: тех вершин сети, которые следуют за анализируемой вершиной. Этот ранг пересчитывается итеративно [Kharlamov, 2020a] с учетом и весов вершин, и весов связей, связанных с исходной вершиной.

Система оказывается эффективной, так как позволяет выявить оглавление корпуса текстов в виде минимального древовидного подграфа семантической сети [там же] всего множества хранимых в системе текстов, где вершина верхнего уровня подграфа представляет собой основную тему, дочерние вершины – второстепенные темы и так далее по иерархии вниз.

Семантическая сеть корпуса текстов может быть разбита путем разрыва слабых связей на подсети, характеризующие предметные подобласти, составляющие предметную область, описываемую текстами корпуса в совокупности.

Отдельные тексты корпуса могут быть отнесены к той или другой предметной области путем сравнения семантической сети текста с сетями предметных областей. Сравнение осуществляется вычислением области пересечения сетей. Конкретный текст может быть отнесен более чем к одной предметной области.

Блок поиска отыскивает в корпусе тексты, имеющие наибольшее пересечение семантических сетей с сетью текста-запроса.

Блок поиска встречного текста отыскивает пару или более текстов, семантические сети которых имеют равные по объему пересечения с семантической сетью текста-запроса, не пересекающиеся друг с другом.

Эта структура дополняется подсистемой автоматического структурирования содержания хранилища – формирования оглавления в виде минимального древовидного подграфа семантической сети [там же] всего множества хранимых в системе текстов.

СЕМАНТИЧЕСКАЯ СЕТЬ ЯЗЫКА

Всё, что говорилось об отдельных текстах, можно сказать и о языке. Язык представлен в текстах, следовательно можно говорить о большой семантической сети всего языка, которая автоматически, в зависимости от содержания отдельных частей, разобьется на подчасти, оглавлением которых будет минимальный древовидный подграф этого большого текста. Эта структура возникнет автоматически, исключительно под воздействием содержания этого большого текста. Так, добавление появляющихся новых текстов не вызывает проблем со структурированием в будущем. И всё это происходит ассоциативно, т. е. исключительно в зависимости от анализируемых текстов, более или менее объективно.

НЕОДНОРОДНАЯ СЕМАНТИЧЕСКАЯ СЕТЬ

Однородность семантической сети в данном случае является следствием двух причин: внутренней и внешней. Во-первых, в мозге человека сеть формируется однородной, так как неоднородность в нее вносится второй частью – экстралингвистическим представлением – это внешняя причина. Но в мозге это сделано умышленно: однородная сеть более робастна, и это вторая, внутренняя причина.

Однако можно сеть сделать и неоднородной. Тогда она размоется на более подробную, но менее представительную структуру, которая будет более детально представлять события мира, но точность вычислений на ней (по причине ее размытости) будет существенно меньше.

ПРОГРАММНАЯ РЕАЛИЗАЦИЯ СИСТЕМЫ

Прототип системы реализован в виде программной технологии TextAnalyst [Kharlamov, 2020b], включающей персональный продукт, а также SDC для встраивания в приложения. Подход достаточно универсален, чтобы позволить анализировать тексты не только на европейских языках, но и на китайском. Продукт очень прост в эксплуатации, а результаты анализа текстов из любых предметных областей легко интерпретируются [Kulikov, Kharlamov, 2020].

Экстралингвистическая модель мира предполагает сложную предобработку входных квазитекстов [Харламов, 2017], по этой причине системы на основе такой обработки для анализа, например, зрительных квазитекстов, пока не реализованы. При этом подход универсален настолько, что на

основе описанных принципов реализована подсистема для анализа генетических сетей, успешно работающая в системе поддержки принятия решений в области онкологии [Kulikov, Kharlamov, 2020].

АППАРАТНАЯ РЕАЛИЗАЦИЯ СИСТЕМЫ

Частично система может быть реализована аппаратно. В зависимости от цели реализации она может быть представлена в виде процессорного ядра с блоком памяти для индивидуального применения, может быть реализована в виде распределенной среды связанных между собой систем вычислительных кластеров для создания большого объема хранилища знаний с ассоциативным доступом, а может быть реализована в виде распределенной среды нейропроцессоров с памятью [Харламов, 2017].

ЗАКЛЮЧЕНИЕ

В статье представлены основные теоретические и практические соображения, архитектура и механизмы системы для хранения текстов (расширяется на квазитексты других модальностей). Коротко описан прототип – программная технология TextAnalyst – анализирующая с использованием представленных принципов и подходов тексты на европейских и китайском языках, формирующая гипертекстовую структуру, включающую в свой состав также его семантическую сеть, описывающую содержание текста, помимо самого текста.

СПИСОК ИСТОЧНИКОВ

1. Информационная система анализа научной деятельности (ИСАНД) в области теории управления / Д. А. Губанов, О. П. Кузнецов, Е. А. Курако, Д. В. Лемтюжникова, Д. А. Новиков, А. Г. Чхартишвили // Проблемы управления. 2024. № 3. С. 42–65.
2. Харламов А. А. Ассоциативная память – среда для формирования пространства знаний. От биологии к приложениям. Дюссельдорф: Palmarium Academic Publishing. 2017.
3. Леонтьева Н. Н. Автоматическое понимание текстов. Системы, модели, ресурсы. М.: ACADEMIA. 2006.
4. Kharlamov A., Pulgun M. (Eds.) Neuroinformatics and Semantic Representations. Theory and Applications. Newcastle upon Tyne: Cambridge Scholars Publishing. 2020.
5. Kharlamov A. A Network N-gram Model of the Text. A Topic Tree of the Text – Minimal Tree Subgraph of the Semantic Network // Neuroinformatics and Semantic Representations. Theory and Applications. Kharlamov A. and Pilgun M. (Eds.). Newcastle upon Tyne: Cambridge Scholars Publishing. 2020a. Pp. 114–126.
6. Kharlamov A. TextAnalyst Technology for Automatic Semantic Analysis of Text // Neuroinformatics and Semantic Representations. Theory and Applications / A. Kharlamov and M. Pilgun (Eds.). Newcastle upon Tyne: Cambridge Scholars Publishing. 2020b. Pp. 156–167.
7. Kulikov A., Kharlamov A. Using a Homogeneous Semantic Network to Classify the Results of Genetic Analysis. In: Neuroinformatics and Semantic Representations. Theory and Applications. A. Kharlamov and M. Pilgun (Eds.). Newcastle upon Tyne: Cambridge Scholars Publishing. 2020. Pp. 219–231.

REFERENCES

1. Gubanov, D. A. et al. (2024). ISAND: an information system for scientific activity analysis (in the field of control theory and its applications) / D. A. Gubanov, O. P. Kuznetsov, E. A, Kurako, D. V. Lemtiuzhnikov, D. A. Novikov, A. G. Tchartishvili. Control Sciences, 3, 42–65. (In Russ.)
2. Kharlamov, A. A. (2017). Assotsiativnaya pamyat' – sreda dlya formirovaniya prostranstva znaniy. Ot biologii k prilozheniyam = Associative memory – a medium for knowledge space formation. From biology to applications. Dusseldorf: Palmarium Academic Publishing. (In Russ.)
3. Leont'eva, N. N. (2006). Avtomaticheskoe ponimanie textov. Sistemy, modely, resursy = Automatic understanding of texts. Systems, models, resources. Moscow: ACADEMIA. (In Russ.)
4. Kharlamov, A., Pilgun, M. (Eds.). (2020). Neuroinformatics and Semantic Representations. Theory and Applications. Cambridge Scholars Publishing.

5. Kharlamov, A. (2020a). A Network N-gram Model of the Text. A Topic Tree of the Text – Minimal Tree Subgraph of the Semantic Network. By A. Kharlamov, M. Pilgun (Eds.) Neuroinformatics and Semantic Representations. Theory and Applications (pp. 114–126). Cambridge Scholars Publishing.
6. Kharlamov, A. (2020b). TextAnalyst Technology for Automatic Semantic Analysis of Text. In Kharlamov, A., Pilgun, M. (Eds.) Neuroinformatics and Semantic Representations. Theory and Applications (pp. 156–167). Cambridge Scholars Publishing.
7. Kulikov, A., Kharlamov, A. (2020). Using a Homogeneous Semantic Network to Classify the Results of Genetic Analysis. In A. Kharlamov, M. Pilgun (Eds.) Neuroinformatics and Semantic Representations. Theory and Applications (pp. 219–237). Cambridge Scholars Publishing.

ИНФОРМАЦИЯ ОБ АВТОРЕ

Харламов Александр Александрович
доктор технических наук
старший научный сотрудник Института высшей нервной деятельности и нейрофизиологии
Российской академии наук

INFORMATION ABOUT THE AUTHOR

Kharlamov Alexander Alexandrovich
Doctor of Technical Science (Dr. habil), Prof.
Senior Researcher at the Institute of Higher Nervous Activity and Neurophysiology
Russian Academy of Sciences

Статья поступила в редакцию	03.07.2024	The article was submitted
одобрена после рецензирования	31.07.2024	approved after reviewing
принята к публикации	06.08.2024	accepted for publication