

TABIIY TILNI QAYTA ISHLASH (NLP)

ПРОЕКТИРОВАНИЕ ГЕОИНФОРМАЦИОННОЙ СИСТЕМЫ "ДИАЛЕКТОЛОГ" С ИСПОЛЬЗОВАНИЕМ НЕЧЕТКОЙ ЛОГИКИ

Галимов Мансур Радикович

Студент по компьютерным наукам, студент

magl.galimov@gmail.com

Институт Прикладной Семиотики Академии Наук Республики Татарстан

Казанский Федеральный Университет, Казань, Россия

Бурнашев Рустам Арифович

Кандидат технических наук, к.т.н

r.burnashev@inbox.ru

Институт Прикладной Семиотики Академии Наук Республики Татарстан

Казанский Федеральный Университет, Казань, Россия

Гатиатуллин Айрат Рафизович

Кандидат технических наук, к.т.н

agat1972@mail.ru

Институт Прикладной Семиотики Академии Наук Республики Татарстан

Казанский Федеральный Университет, Казань, Россия

Прокопьев Николай Аркадиевич

Магистр по компьютерным наукам, магистр

nikolai.prokopyev@gmail.com

Институт Прикладной Семиотики Академии Наук Республики Татарстан

Казанский Федеральный Университет, Казань, Россия

Аннотация. В данном докладе рассматриваются результаты научных исследований по проектированию прототипа нечеткой экспертной системы для обработки знаний в области диалектологии, а также последующей визуализации данных с использованием геоинформационных систем и технологий. Своими исследованиями мы стремимся внести свой вклад в развитие интеллектуальных языковых информационных ресурсов с использованием геоинформационных систем и технологий. Для разработки веб-интерфейса программной оболочки был использован фреймворк Django языка программирования Python. Для анализа, обработки и визуализации данных использовались программные библиотеки Pandas и Folium. Для повышения эффективности обработки знаний была использована графовая база данных neo4j.

Ключевые слова: нечеткие экспертные системы, база знаний, Folium, Pandas, ГИС.

Annotation. This report discusses the results of scientific research on designing a prototype fuzzy expert system for processing knowledge in the field of dialectology, as well as subsequent data visualization using geographic information systems and technologies. With our research, we strive to contribute to the development of intelligent language information resources using geographic

information systems and technologies. The Django framework of the Python programming language was used to develop the shell web interface. Pandas and Folium software libraries were used for data analysis, processing and visualization. To increase the efficiency of knowledge processing, the neo4j graph database was used.

Keywords: fuzzy expert systems, knowledge base, Folium, Pandas, GIS.

ВВЕДЕНИЕ

В настоящее время диалектолог взаимодействует с большим набором языковых информационных ресурсов и баз данных. Из-за большого объема данных специалисту трудно найти необходимую информацию в процессе профессиональной деятельности. В связи с этим одним из главных приоритетов является создание современных программных приложений с элементами искусственного интеллекта, связанных геопространственными данными.

В профессиональной деятельности диалектолога возникает трудность в сборе и обработке входной информации, получаемой в режиме реального времени. Использование на практике современных средств геоинформационных технологий для считывания и обработки различий диалектов народов с последующим картографированием является актуальной задачей. Использование лингвогеографического метода повысит эффективность принятия решений во время научных экспедиций и исследований.

Мы также провели эксперимент по внедрению графической базы данных neo4j для работы с геоинформационной системой.

В качестве одной из последних разработок в области создания лингвистической ГИС можно выделить работу Дж. Мороза из Лаборатории языковой конвергенции и Школы лингвистики, программа *lingtypology* [1]. Эта программа представляет собой бесплатный R-пакет для картографирования и лингвистической типологии.

Лингвистическое картографирование [3] является относительно устоявшейся областью исследований в лингвистике, но только недавно лингвогеография начала использовать геоинформационные системы (ГИС) в качестве инструмента [4]. Сочетание географических характеристик с типологическими исследованиями и новейшими достижениями в области информационных технологий позволяет создавать современные инструменты, такие как автоматизированное рабочее место диалектолога, типолога, этимолога и т.д. Использование таких инструментов позволяет осуществлять поиск корреляций между языковыми параметрами и географическим распределением языков. Инструментарий расширяется за счет использования статистических технологий, что, например, позволяет усовершенствовать методы оценки географических расстояний между различными языковыми сообществами. Несмотря на появление работ, использующих методы

лингвогеографии, не существует единого технологического инструментария, в частности, для интеграции баз данных предыдущих исследований, и это приводит к тому, что каждая созданная карта создается один раз для определенных целей и не может быть использована для других исследований.

Основой для создания диалектов разных языков являются теоретические положения лингвистической географии [3]. Картографирование диалектных явлений основано на следующих концепциях:

- диалектный язык - это совокупность диалектов языка, включающая как черты, общие для всего языка, так и черты, которые имеют различные варианты в диалектах, но в то же время являются элементами общей языковой системы.;
- диалектное различие, формирующее соответствующее явление (в более поздних работах также междиалектное соответствие), является основным объектом отображения в атласе, лингвистическим явлением, которое в диалектах проявляется в различных вариантах. Graphic information is perceived several times faster [4] than text information.

В настоящее время ГИС-технологии активно используются системами искусственного интеллекта, в частности, экспертными системами [5-8]. Например:

- MAPEX [9, 10] представляет собой экспертную систему, основанную на правилах автоматического обобщения картографических ресурсов.
- AUTONAP [9, 10] представляет собой экспертную систему для автоматического размещения названий объектов на географической карте. Эта система имитирует действия опытного картографа при выполнении задачи по нанесению названий объектов на карту. AUTONAP использует эвристические знания о размещении имен на основе установленных процедур и соглашений в базе знаний.
- CES [11, 12] является прототипом картографической экспертной системы и используется в качестве картографического консультанта в области энергетики, шахт и т.д. Используя таблицы принятия решений, он предоставляет набор картографических спецификаций, которые оптимально соответствуют набору требований к карте, предъявляемых в системе.

Для разработки нашего программного приложения, использовались библиотеки языка программирования Python: Django, Pandas, Folium, os, NumPy, Neomodel, и т.д.

Библиотека Folium использовалась для работы с интерактивными картами [13].

Folium это библиотека в Python, которая помогает разрабатывать карты Leaflet.js (JavaScript язык программирования). Позволяет нам выполнять обработку данных в Python и визуализировать свои объекты на карте.

Folium позволяет легко визуализировать данные, которые были обработаны в Python на интерактивной карте Leaflet. Он позволяет отображать данные на карте, а также передавать векторные / растровые / HTML-элементы в качестве маркеров на карте.

В библиотеке есть несколько встроенных наборов карт из OpenStreetMap, Mapbox и Stamen (Рис. 1.). Библиотека поддерживает обработку файлов в формате представления различных географических структур данных (GeoJSON и TopoJSON), изображений и видео.



Fig. 1. Добавление географических координат на карту.

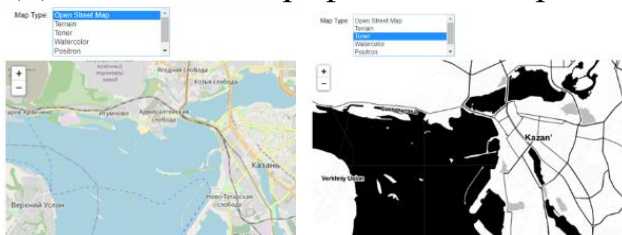


Fig. 2. Добавление интерактивных стилей (OpenStreetMap, Toner)

Обработка и анализ данных. Современные библиотеки программного обеспечения вносят значительный вклад в разработку программных средств и средств обработки знаний. Одним из наиболее важных свойств таких инструментов является общая доступность и полезность их использования. Такие инструменты могут решать интеллектуальные задачи, характерные для людей – классифицировать различные объекты, распознавать закономерности в изображениях, предсказывать события и т.д. Даже сейчас интеллектуальные информационные системы обрабатывают данные и решают проблемы гораздо быстрее и точнее, чем люди.

Для обработки и анализа данных использовалась библиотека pandas [14]. Pandas - это программная библиотека Python для обработки и анализа данных. Функциональность программной библиотеки pandas с данными построена поверх библиотеки NumPy, которая является инструментом более низкого уровня. Предоставляет специальные структуры данных и операции для работы с числовыми таблицами, временными рядами и др.

Библиотека включает в себя объект DataFrame (двумерный массив), аналогичный электронной таблице Excel (данные из файла можно загрузить с

помощью команды `pandas read_csv` ('имя файла')).

С помощью библиотеки программного обеспечения вы можете выполнять те же манипуляции с данными: объединять список, сортировать по требуемому атрибуту, выполнять анализ, генерировать отчеты и т.д.

Ниже (Рис. 3) приведен пример данных, полученных из файла для последующей интеграции в географическую информационную систему диалектолога.

```
In [322]: df
```

```
Out[322]:
```

	word	latitude	longitude	country	city	icon_num
0	жан	55.78	49.13	Russia	Kazan	1
1	жылы	55.90	49.13	Russia	Kazan	2
2	жирне	55.95	49.13	Russia	Kazan	3
3	жухаз	55.78	48.40	Russia	Innopolis	4
4	божра	56.40	49.13	Russia	Kazan	5
5	жөмлө	55.12	49.13	Russia	Kazan	6

Рис. 3. Фрагмент DataFrame после считывания и обработки геопространственных данных.

Ниже приведен программный код для интеграции данных с файловой системы на графическую карту программной библиотеки Folium (Рис. 4.):

```
In [313]: # plot dialects locations
for (index, row) in df.iterrows():
    folium.Marker(location=[row.loc['latitude'], row.loc['longitude']],
                  popup=row.loc['word'] + ' ' + row.loc['city'] + ' ' + row.loc['country'],
                  tooltip='click').add_to(folium_map)

# display map
folium_map
```

Рис. 4. Фрагмент кода для добавления данных на карту

Для работы с графиками и отчетами использовалась программная библиотека Matplotlib. Matplotlib — это библиотека на языке программирования Python для визуализации данных с помощью двумерной (2D) графики. Обработанные графики могут быть использованы в качестве иллюстраций при формировании отчета (статистики).

Формат файла GeoJSON предназначен для хранения географовых структур данных, основанных на JSON.

GeoJSON поддерживает следующие типы объектов:

- Точка (включая адреса и местоположения).
- Строка линий (включая улицы, шоссе, границы).
- Полигон (включая страны, провинции и земельные участки).
- Составные объекты типов point, linestring или polygon.

Ниже приведен листинг программного кода файла GeoJSON:

```
{
  "type": "FeatureCollection",
  "features": [
    {
      "type": "Feature",
      "properties": {},
      "geometry": {
        "type": "Polygon",
        "coordinates": [
          [[46.120605, 57.01192],
            [48.537598, 57.569959],
            [51.61377, 56.856072],
```

[52.536621,54.782834],
.....,
[44.736328,55.110943],
[46.120605,57.01192]]]]}}}}

Ниже приведен прототип системы (Рис. 4.), созданный в результате исследований по обработке знаний и последующей визуализации географических структур данных.

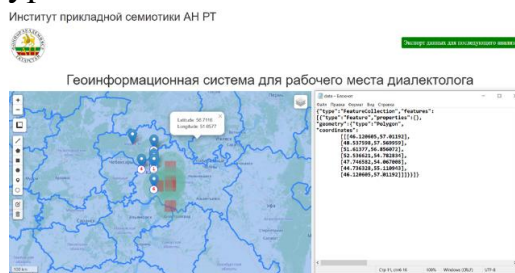


Рис. 4. Основная форма с интерактивными элементами на карте

Программная библиотека Neomodel была использована при разработке для реализации следующей архитектуры. Каждый язык может иметь связи с регионами (Рис. 5.), в свою очередь, язык может содержать слова, которые могут использоваться сразу в нескольких диалектах.

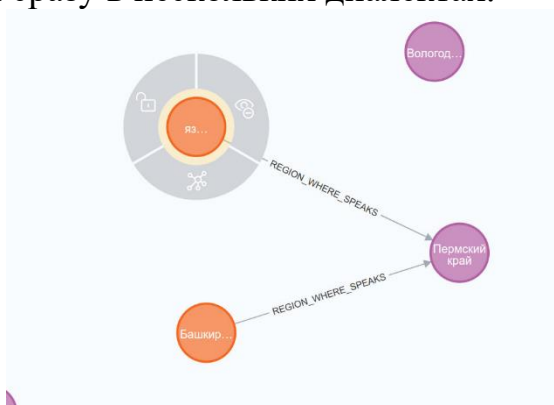


Рис. 5. Пример взаимосвязи между регионом и диалектом

Таким образом, мы получаем набор узлов (Рис. 6.), которые можно выводить в геоинформационную систему и взаимодействовать с ней для получения дополнительной информации об узле.

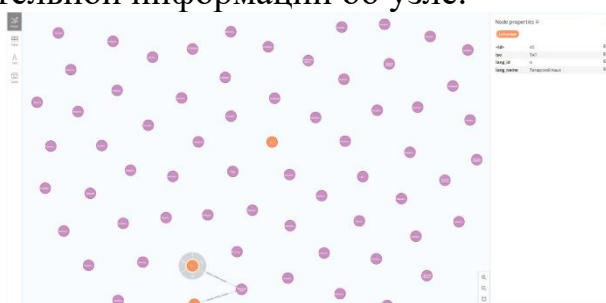


Рис. 6. Набор регионов и языковых узлов.

В качестве примера разработанной модели возьмём Language. Модель Neo позволяет связать модели Django со спецификой Neo4j, где соединения раскрываются с помощью необычного внешнего ключа, а также RelationshipTo и RelationshipFrom.

```
Class Language(StructuredNode,NodeUtils:  
    lang_code = IntegerProperty()  
    iso = StringProperty(max_length=5)  
    lang_id=IntegerProperty(index=True)  
    lang_name = StringProperty(max_length=100, required=True)  
    regions = RelationshipTo(".region.Region",  
"REGION_WHERE_SPEAKS")
```

Экспертные системы - это информационные системы, которые частично или полностью заменяют эксперта в определенной предметной области. Основной целью ЭС является разработка программных средств, которые при решении сложных для человека задач получают результаты, не уступающие по качеству и эффективности решениям, получаемым человеком-экспертом. Главным преимуществом экспертных систем является способность накапливать знания, сохранять их в течение длительного времени, обновлять и тем самым обеспечивать относительную независимость конкретной организации от наличия в ней квалифицированных сотрудников. Накопление знаний позволяет повышать квалификацию специалистов, работающих на предприятиях, используя лучшие, проверенные решения. Практическое применение систем искусственного интеллекта в организациях и в экономике основано на экспертных системах, которые повышают качество и сокращают время принятия решений, а также способствуют увеличению объема работы и профессиональному развитию специалистов.

Основной группой методов, используемых в нечеткой экспертной системе (ЭС) [15-17], предлагаемой в проекте, являются методы нечеткой логики. Теория нечетких множеств была предложена американским математиком Лотфи Заде [18]. Применение теории нечетких множеств позволяет описывать нечеткие понятия и знания, оперировать этими знаниями и делать нечеткие выводы. Неопределенные данные и ситуации обрабатываются с использованием нечеткой логики с использованием лингвистических переменных [19, 20].

Нечеткая логика – это логическая система или система управления n-разрядной логической системой, которая использует степени состояния ("степени истинности") входных данных и генерирует выходные данные, которые зависят от состояний входных данных и скорости изменения этих состояний.

Нечеткая операция предполагает использование нечетких множеств и функций принадлежности. Каждый нечеткий набор является представлением лингвистической переменной, которая определяет возможное состояние

выходного сигнала. Функция принадлежности - это функция общего значения в нечетком множестве, так что и общее значение, и нечеткое множество принадлежат универсальному множеству.

Степени принадлежности к этому общему значению в нечетком множестве определяют выходные данные на основе принципа "Если-То". Принадлежность присваивается на основе предположения о выходе с использованием входных данных и скорости изменения входных данных. Функция принадлежности - это, по сути, графическое представление нечеткого множества. Структура нечеткой экспертной системы показана ниже (Рис. 7).

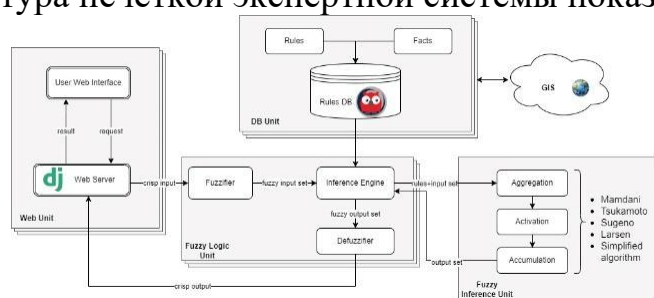


Рис. 7. Структура прототипа нечеткой экспертной системы

База знаний нечеткой экспертной системы включает в себя полную информацию о данных исследований в области диалектологии тюркских языков. Экспертная система включает в себя следующие программные модули: модуль сбора знаний, решатель, пояснительный и диалоговый компоненты.

Модуль получения знаний автоматизирует процесс наполнения новыми знаниями с использованием геоинформационных технологий (ГИС).

Решатель, используя исходные данные из рабочей памяти и знания из баз знаний, формирует последовательность правил логического вывода, которые впоследствии приводят к решению задач в области диалектологии.

Пояснительный и диалоговый компоненты предназначены для получения количественных и качественных показателей для выбранного на карте диалектологического атласа и организации дружественного интерфейса, что повышает эффективность консультаций в области диалектологии.

Логическая программа состоит из предикатов, которые представляют собой функции, генерирующие логические значения. База знаний формируется с использованием логического слова-сказуемого, которое включает в себя факты о предметной области.

Таким образом, разработка нечеткой экспертной системы будет осуществляться на основе представленных этапов разработки, с промежуточной проверкой используемых методов нечеткой логики и лингвистических переменных для конкретных идиом (языков и диалектов). Интеллектуальный интерфейс ЭС позволит специалистам по тюркским языкам и диалектам работать при принятии решений в условиях неполных и противоречивых данных.

База знаний нечеткой экспертной системы будет взаимосвязана с научной работой [21], которая проводится в Институте прикладной семиотики Академии наук Республики Татарстан. Инструменты этого портала позволяют вам описывать как языки, так и диалекты на разных лингвистических уровнях и в разных проявлениях.

Акцент на диалектах делается на том, что разные диалекты одного и того же языка могут быть представлены в базе данных портала как отдельные языки. А такие концептуальные семантические структуры, как тезаурусы типа WordNet [22] и Frameset frame base, позволяют описывать семантику языковых единиц в разных диалектах и проводить сравнительные исследования между диалектами.

Благодарность

Исследование выполнено при финансовой поддержке Республики Татарстан в рамках проекта молодежного научного гранта «Разработка программного приложения для визуализации данных о диалектах с использованием геоинформационных технологий»

Заключение

С использованием современных средств визуализации данных удастся представить информацию максимально доступной для восприятия специалистом в предметной области.

Для интеллектуализации системы планируется дальнейшая интеграция экспертной системы в оболочку геоинформационной системы.

Разработанный проект планируется в будущем интегрировать в структуру портала "Турецкая морфема". ([http://modmorph.turklang.net /](http://modmorph.turklang.net/)).

Для разработки веб-интерфейса программной оболочки был использован фреймворк Django языка программирования Python, для анализа, обработки и визуализации данных использовались программные библиотеки Pandas и Folium.

Список литературы:

1. lingtypology: introduction and installation, Electronic resource: ropensci.github.io/lingtypology
2. Egorov, I.M., Dybo, A.V. and Kassian, A.S., 2022. Phylogeny of the Turkic Languages Inferred from Basic Vocabulary: Limitations of the Lexicostatistical Methods in an Intensive Contact Situation. *Journal of Language Evolution*, 7(1), pp.16-39.
3. Salimov F. I. i dr. Jelektronnaja versija atlasa tatarskih narodnyh govorov // *Filologija i kul'tura*. – 2011. – №. 26. – S. 205-210.
4. Zhdanova E. A., Belyh A. A. Geograficheskie informacionnye sistemy v lingvisticheskix issledovaniyax // *Intellectual'nye sistemy v proizvodstve*. Izhevsk, 2014. № 2 (24). S. 169–174.
5. Frazer, T.C., Breton, R.L., & Schiffman, H.F. (1991). *Geolinguistics: Language Dynamics and Ethnolinguistic Geography*.

6. Briscoe, Ulla. (2009). Geolinguistic GIS Applications: Aspect of Data Quality in Mapping Lesser-Used Languages, Centre for Geoinformatics, Salzburg University
7. R. Gibadullin and N. Marushkai, "Development of Predictive CNN Based Model for Vital Signs Alerts," in International Conference on Industrial Engineering, Applications and Manufacturing (ICIEAM), 2021, pp. 404-409.
8. R. Burnashev, I. Enikeev and A. Enikeev, "Design and Implementation of Integrated Development Environment for Building Rule-Based Expert Systems," 2020 International Multi-Conference on Industrial Engineering and Modern Technologies (FarEastCon), 2020, pp. 1-4.
9. Choi, Jinmu & Usery, E.. (2004). System Integration of GIS and a Rule-Based Expert System for Urban Mapping. Photogrammetric Engineering & Remote Sensing. 70. 217-224.
10. Eldrandaly, Khalid. (2008). Integrating Expert Systems and GIS for Spatial Decision Making: Current Practices and New Trends.
11. Fisher, P. F. (1989). Expert System Applications in Geography. Area, 21(3), 279–287.
12. Rodriguez-Bachiller, Agustin & Glasson, John. (2004). Expert Systems and Geographic Information Systems for Impact Assessment.
13. K.A. Fedutinov. Structuring environmental information using geographic information technologies. Modeling, optimization and information technology. 2019;7(4). Available from: https://moit.vivt.ru/wp-content/uploads/2019/11/Fedutinov_4_19_1.pdf DOI: 10.26102/2310-6018/2019.27.4.044
14. Il'ichev V.Ju., Jurik E.A. Analiz massivov dannyh s ispol'zovaniem biblioteki Pandas dlja Python // Nauchnoe obozrenie. Tehnicheskie nauki. – 2020. – № 4. – S. 41-45
15. Pospelov D.A. (Red.). Jekspertnye sistemy: sostojanie i perspektivy. 1989. 152 s.
16. Popov Je. V. Osobennosti razrabotki i ispol'zovaniya jekspertnyh sistem //Iskusstvennyj intellekt. Kn. – 1990. – T. 1.
17. Averkin A. Hybrid intelligent systems based on fuzzy logic and deep learning //Artificial Intelligence. – Springer, Cham, 2019. – S. 3-12.
18. Zadeh, L.A., 1988. Fuzzy logic. Computer, 21(4), pp.83-93
19. Zadeh L.A. From computing with numbers to computing with words – from manipulation of measurement to manipulation of perceptions – IEEE Trans. On Circuits and Systems -1: Fundamental Theory and Applications, 45, 1, 1999, 105–119.
20. Demenkov N.P. Jazyk nechetkogo upravleniya // Promyshlennyye ASU i kontrollery. – 2005. – № 5. – S. 30–36.
21. Gatiatullin A., Suleymanov D., Prokopyev N., Khakimov B. (2020) About Turkic Morpheme Portal. CEUR Workshop Proceedings Institute for history,



language and literature, Ufa scientific center, Russian Academy of Sciences Proceedings of TurkLang 2020, pp. 226-243.

22. Gatiatullin, A., Kubedinova, L. Multilingual thesaurus for turkic languages (2020) 5th International Conference on Computer Science and Engineering, UBMK 2020, 2020, pp. 393–398, 9219378