DOI: 10.55648/1998-6920-2024-18-1-70-77 УДК 004.8:377

Лингвистическое обеспечение информационных систем и процессов

И. В. Перцев, Е. И. Ситняковская

Сибирский гос. унив. телекоммуникаций и информатики (СибГУТИ)

Аннотация: Статья рассматривает процесс проектирования словарей данных, которые представляют собой наборы лексических единиц с определенной информацией о них. Описываются методы создания словарей индексирования и поиска информации, которые позволяют эффективно организовывать поиск по текстовым данным с использованием специализированных индексов.

Ключевые слова: лингвистическое обеспечение, информационные системы, словари, индексирование, поиск информации, тезаурусы, лексические комплексы, семантический анализ, синтаксический анализ, базы данных.

Для цитирования: Перцев И. В., Ситняковская Е. И. Лингвистическое обеспечение информационных систем и процессов // Вестник СибГУТИ. 2024. Т. 18, № 1. С. 70–77. https://doi.org/10.55648/1998-6920-2024-18-1-70-77.

Контент доступен под лицензией Creative Commons Attribution 4.0 License

© Перцев И. В., Ситняковская Е. И.

Статья поступила в редакцию 15.12.2023;
принята к публикации 10.01.2024.

1. Введение

Лингвистическое обеспечение информационных систем играет важную роль в разработке и эффективном функционировании таких систем. Оно позволяет улучшить качество и эффективность работы с текстами, обеспечивая точный и быстрый поиск информации, а также удобное взаимодействие пользователей с системой.

Понятие «лингвистическое обеспечение» включает в себя различные аспекты и задачи, связанные с использованием языка и текстов в информационных системах. Оно заключается в применении лингвистических методов анализа, классификации и интерпретации текстов, а также в разработке соответствующих инструментальных средств, таких как словари данных, словари индексирования и поиска информации, тезаурусы и другие лексические комплексы [1].

Проектирование словарей индексирования и поиска данных является ключевым инструментом в обеспечении быстрого и точного поиска информации. Будут рассмотрены инвертированные списки терминов, деревья поиска и семантический анализ для создания эффективных словарей и индексов для поиска информации.

Тезаурусы и другие лексические комплексы используются для структурирования и классификации информации. Также будет исследовано применение методов семантического, синтаксического и прагматического анализа текстовой информации для представления данных в базах данных и организации интерфейсов информационных систем с пользователями [2].

Так как в современном информационном обществе огромное количество данных представлено в текстовой форме, эффективное управление и поиск этой информации являются важной задачей. Лингвистическое обеспечение позволяет осуществлять адекватный анализ текстов, обнаруживать и извлекать смысловую информацию, а также предоставлять удобный доступ к данным. Кроме того, лингвистическое обеспечение способствует улучшению качества информационных систем и процессов. Оно позволяет автоматизировать задачи, связанные с обработкой текстов, автоматически разбирать и классифицировать информацию, а также создавать удобные интерфейсы для взаимодействия пользователей с системой. Всё это способствует повышению эффективности работы с текстовой информацией и улучшению результатов в различных областях, таких как поиск информации, анализ текста, машинный перевод, автоматическая обработка естественного языка и другие.

Таким образом, лингвистическое обеспечение информационных систем играет важную роль в поиске, обработке и управлении текстовой информацией. Его использование способствует улучшению качества работы систем и повышению эффективности процессов.

2. Методы проектирования словарей

Методы проектирования словарей данных играют важную роль в обеспечении эффективной организации и управления данными в базах данных.

Первоначальный шаг в проектировании словарей данных – определение требований и понимание предметной области, с которой будет работать словарь данных. Это позволит определить основные термины и связи между ними [3]. Словарь данных может быть организован в виде иерархической структуры, где различные термины и их связи могут быть представлены в виде древовидной структуры. Также можно использовать другие методы организации, такие как сети или списки терминов с определениями и атрибутами. Для каждого термина в словаре данных необходимо предоставить четкое и однозначное определение. Определение должно быть достаточно подробным, чтобы пользователь мог понять значение и использование термина, а также связи с другими терминами. Важно вести документацию словаря данных, чтобы обеспечить ее актуальность и доступность для пользователей. Также можно предусмотреть механизмы обновления и внесения изменений в словарь данных при необходимости. Существуют различные средства, которые могут быть использованы для проектирования и поддержки словарей данных. Например, специализированные программы или инструменты баз данных могут предоставлять функциональность для определения и организации терминов, а также для отслеживания изменений.

Проектирование словарей включает определение требований, создание структуры и организацию терминов, описание терминов, документирование и поддержку словаря данных, а также использование средств автоматизации. Определение требований и анализ предметной области позволяют определить основные термины и связи между ними. Создание структуры и организация терминов способствуют логичной и понятной систематизации данных. Описание терминов обеспечивает единообразное понимание и использование терминологии в системе. Документирование и поддержка словаря данных помогают поддерживать актуальность и качество словаря. Использование средств автоматизации, таких как специальные программы или инструменты баз данных, повышает эффективность процесса проектирования словарей данных.

В целом методы проектирования словарей данных являются неотъемлемой составляющей обеспечения эффективной организации и управления данными в информационных системах. Качественно проведенное проектирование словарей данных позволяет улучшить эффективность работы с данными в информационных системах.

Примеры методов и средств проектирования словарей данных:

1. Создание табличных словарей. В этом методе каждый термин и его определение хранятся в отдельной записи таблицы с соответствующими полями. Такой подход удобен для поиска и управления данными.

2. Использование онтологий и тезаурусов. Онтологии и тезаурусы позволяют описать термины и их отношения между собой. Это помогает более полно и точно описать данные и взаимосвязи между ними [4].

3. Иерархические словари. Этот метод предполагает организацию терминов в иерархическую структуру, где каждый термин имеет родительский и дочерний элементы. Такая организация позволяет более удобно классифицировать данные и проводить поиск по иерархии.

4. Семантические словари. Семантический словарь предоставляет информацию о значениях и смысле терминов. Он описывает лингвистические и семантические свойства слова, его синтаксические части, семантические роли и другие связи. Семантические словари помогают улучшить поиск и анализ данных.

5. Использование графических средств. Графические средства, такие как схемы и диаграммы, могут использоваться для визуализации словарей данных. Это позволяет лучше представить связи и ассоциации между терминами и сущностями, а также облегчить понимание данных.

Хороший словарь данных должен иметь четкие и универсальные определения терминов, чтобы избежать путаницы и неоднозначности. Словари данных должны учитывать особенности предметной области и контекст, в котором будут использоваться данные. Должны быть определены специфические термины и понятия, которые относятся к данной области. Словари данных должны быть гибкими и способными к дальнейшему расширению, чтобы удовлетворить потребности в новых терминах и понятиях, должны быть согласованы и сопоставимы с другими используемыми словарями и схемами данных, чтобы обеспечить единообразное понимание и использование данных. Словари данных должны быть легкодоступными и удобными в использовании, чтобы пользователи могли быстро находить нужную информацию и применять ее при работе с данными.

Таким образом, методы и средства проектирования словарей данных предоставляют возможности для эффективного хранения, организации и поиска информации. Соблюдение принципов проектирования словарей данных позволяет создать удобные и надежные инструменты для работы с текстовыми данными.

3. Этапы работы системы лингвистического обеспечения

Система работы лингвистического обеспечения информационных систем и процессов включает несколько этапов:

1. Определение требований и анализ предметной области: на этом этапе осуществляется изучение конкретной предметной области и выявление основных терминов и связей между ними.

2. Проектирование словарей данных: создаются словари, которые содержат определения терминов, связи между ними и другую сопутствующую информацию.

3. Создание словарей индексирования и поиска информации: разрабатываются специализированные словари и индексы, которые позволяют эффективно организовывать поиск и индексацию текстовой информации.

4. Применение методов семантического, синтаксического и прагматического анализа: эти методы используются для обработки текстовой информации и ее представления в базах данных. Семантический анализ позволяет выявить смысловую информацию, синтаксический анализ занимается разбором грамматической структуры предложений, а прагматический анализ учитывает контекст и цели коммуникации.

5. Организация интерфейсов информационных систем с пользователями: разрабатываются удобные и понятные интерфейсы, обеспечивающие эффективное взаимодействие пользователя с информационной системой.

Таким образом, система работы лингвистического обеспечения информационных систем и процессов включает этапы определения требований и анализа предметной области, проектирования словарей данных, создания словарей индексирования и поиска, применения методов анализа текстовой информации, а также организации интерфейсов с пользователями [5]. Каждый этап имеет свою значимость и важность для эффективной работы информационной системы.

Словари индексирования и поиска информации являются важной составляющей информационных систем, так как они позволяют обеспечить быстрый и точный поиск данных. Они упрощают обработку и анализ больших объемов информации, позволяют организовать эффективное хранение данных и предоставляют возможность пользователям быстро находить нужную информацию. Словарь индексирования – это специальная структура данных, используемая для организации и упорядочивания информации для эффективного поиска. Он содержит отображение между ключевыми словами или терминами и связанными с ними документами или записями.

4. Средства индексирования и поиска информации

Приведём некоторые примеры методов и средств проектирования словарей индексирования и поиска информации.

Метод инвертированного списка терминов. Это один из наиболее популярных методов, используемых для построения словарей индексирования и поиска. Здесь каждый термин или ключевое слово связывается с соответствующими документами или записями. Это позволяет быстро находить необходимые данные по ключевым словам.

Деревья поиска. Это структуры данных, используемые для организации и упорядочивания данных. Они позволяют эффективно искать данные в больших объемах информации. Примерами деревьев поиска являются бинарные деревья поиска, или B-деревья.

Методы семантического анализа. Семантический анализ позволяет учитывать смысловую информацию при индексировании и поиске данных. Он помогает улучшить качество поиска и точность результатов поиска. Примеры методов семантического анализа включают индексирование по семантическим полям или использование онтологий.

Тезаурус является специальным словарем, который содержит систематизированное описание и организацию концептуальных связей между терминами или понятиями в определенной предметной области. Лексический комплекс, с другой стороны, представляет собой набор лексических данных, включающих разнообразные смыслы, значения, синтаксические свойства и другие характеристики слов или фраз.

Тезаурусы и лексические комплексы играют важную роль в информационных системах, они обеспечивают точность и полноту поиска информации. Введение тезаурусов и лексических комплексов позволяет учесть синонимы, близкие по значению термины и альтернативные формулировки, что помогает расширить охват поиска и улучшить точность результатов, обеспечить поддержку классификации и организации данных.

Тезаурусы и лексические комплексы могут использоваться для структурирования, классификации и организации информации в информационных системах, создавая иерархические или сетевые связи между связанными терминами. Тезаурусы и лексические комплексы помогают понимать и интерпретировать запросы пользователей, что позволяет более точно и полно находить необходимую информацию. К примеру, семантический анализ позволяет идентифицировать отношения между терминами и определить их семантические характеристики. Это могут быть связи типа «является», «часть – целое», «синонимы» и другие.

Методы семантического анализа могут включать использование лингвистических моделей, машинного обучения и статистических методов. Одним из методов проектирования тезаурусов является использование иерархических отношений, где термины организованы в структуру с родительскими и дочерними элементами. Это позволяет классифицировать термины по определенным критериям и создавать иерархическую структуру понятий. Также можно рассмотреть онтологическое моделирование. Онтологии представляют собой формальные описания понятий и связей в предметной области. Создание онтологий позволяет формализовать семантические отношения между терминами и обеспечить единообразное понимание. Применение онтологического моделирования включает использование RDF (расширяемый язык разметки ресурсов), OWL (язык описания онтологий) и других стандартов [6]. Семантические сети и графы могут быть применены для визуализации и организации тезаурусов и лексических комплексов. Они позволяют визуально представить связи между терминами и концептуальные отношения, упрощая восприятие и понимание данных.

Семантический анализ – это процесс извлечения и интерпретации смысловой информации из текстов. В контексте обработки текста семантический анализ означает понимание значения слов, синтаксиса и семантических отношений между ними.

Семантический анализ может быть использован для представления текстовой информации в базах данных. Он позволяет обработать и структурировать тексты таким образом, чтобы они стали доступными для поиска и извлечения нужной информации из базы данных. Семантический анализ обеспечивает более точное сопоставление запросов пользователей с содержанием текстовых данных в базе. Например, ключевые слова и извлечение сущностей. Методы извлечения ключевых слов и сущностей позволяют выделить наиболее значимые слова или фразы из текста, которые могут служить ключевыми понятиями для индексации и поиска данных [7].

5. Организация интерфейсов информационных систем

Организация интерфейсов информационных систем играет важную роль в обеспечении удобного и эффективного взаимодействия пользователя с системой [8]. Интерфейсы являются способом представления функциональности и данных системы, а также механизмом взаимодействия и обмена информацией с пользователем. Хорошо организованные интерфейсы делают работу с системой более удобной и понятной для пользователей, что приводит к повышению эффективности и производительности. Четко организованный и интуитивно понятный интерфейс снижает возможность ошибок пользователей и помогает предотвратить их возникновение, что повышает надежность всей системы. Хорошо спроектированный интерфейс позволяет пользователям быстро освоить функциональность системы и легко адаптироваться к ней, уменьшая время и усилия, необходимые для обучения. Грамотно организованный интерфейс учитывает потребности, навыки и предпочтения пользователей, что повышает удовлетворенность пользователей системой [9].

Предварительное исследование пользовательского опыта, проведение опросов и наблюдений помогает понять потребности и предпочтения пользователей, что является основой для организации интерфейса. Интерфейсы должны быть интуитивно понятными и легкими для использования, чтобы пользователи могли без лишних усилий выполнять необходимые задачи. В качестве методов можно использовать принципы дизайна, такие как простота, согласованность, видимость и гибкость. Существуют стандарты и руководства по проектированию интерфейсов, которые содержат рекомендации и принципы для создания удобных и эффективных интерфейсов, такие как стандарты ISO (например, ISO 9241-110:2018 "Human-centred design for interactive systems") и руководства по дизайну пользовательских интерфейсов [10].

Проведение тестирования интерфейса с участием пользователей позволяет выявить проблемы и улучшить его дальнейшими итерациями. Важно получать обратную связь пользователей и активно учитывать ее при переработке интерфейса.

6. Заключение

Дальнейшие исследования могут быть направлены на разработку новых методов и средств проектирования словарей данных, словарей индексирования и поиска информации, тезаурусов и лексических комплексов, чтобы обеспечить еще более эффективную и удобную работу с текстовой информацией. Важным направлением исследований является разработка улучшенных методов семантического анализа текстовой информации, которые позволят более точно и полно понимать смысл и контекст текстов и улучшить результаты поиска и анализа.

В совокупности лингвистическое обеспечение информационных систем играет важную роль в обработке и управлении текстовой информацией. Оно основано на методах и средствах проектирования словарей данных, словарей индексирования и поиска информации, тезаурусов и лексических комплексов, а также на методах семантического, синтаксического и прагматического анализа текстовой информации. Вместе они помогают повысить качество работы информационных систем и обеспечить удобство взаимодействия пользователей с системой. Важно также провести дальнейшие исследования в области эффективного использования лингвистического обеспечения в конкретных областях, таких как медицина, право, наука и т.д., чтобы разработать специализированные методы и инструменты для этих областей.

Литература

1. *Aitchison J., Gilchrist A., Bawden D. Thesaurus Construction and Use: A Practical Manual*. Aslib-IMI, 2000. 240 p.
2. *Beale S.* An open-source terminology and knowledge representation system: Ohsumed thesaurus design and implementation // Computers and Biomedical Research. 1996. V. 29, Is. 3. P. 193–223.
3. *Hodge G.* Systems of thesaurus compilation: a pilot study of thesaurus design and construction // Journal of Documentation. 2000. V. 56 (2). P. 145–168.
4. *Shum S. B., Motta E., Domingue J.* Ontology-based knowledge representation for aerospace design // Journal of Knowledge Engineering and Knowledge Management. 2001. V. 19 (5). P. 514–536.
5. *Jurafsky D., Martin J. H.* Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition (3rd Edition). Pearson, 2023. 577 p.
6. ISO 25964-1:2011 Thesauri and interoperability with other vocabularies. Part 1: Thesauri for information retrieval.
7. *Studer R., Benjamins V. R.* Knowledge engineering: Principles and methods // Data & Knowledge Engineering. 1998. V. 25 (1–2). P. 161–197.
8. *Shneiderman B., Plaisant C.* Designing the User Interface: Strategies for Effective Human-Computer Interaction. 5th Ed. Pearson, 2020. 677 p.
9. *Dix A., Finlay J., Abowd G., Beale R.* Human-Computer Interaction. 3rd Ed. Prentice Hall, 2004. 52 p.
10. ISO 9241-110:2018 Ergonomics of human-system interaction. Part 110: Interaction principles.

Перцев Игорь Владимирович

к.т.н., доцент, доцент кафедры прикладной математики и кибернетики, Сибирский государственный университет телекоммуникаций и информатики (СибГУТИ, 630102, Новосибирск, ул. Кирова, д. 86), тел. +7 383 2698 272, e-mail: pertsev@sibguti.ru, ORCID ID: 0009-0005-2069-3937.

Ситняковская Елена Игоревна

к.т.н., доцент, доцент кафедры прикладной математики и кибернетики, Сибирский государственный университет телекоммуникаций и информатики (СибГУТИ, 630102, Новосибирск, ул. Кирова, д. 86), тел. +7 383 2698 272, e-mail: sitlen@sibguti.ru, ORCID ID: 0009-0000-0964-787X.

Авторы прочитали и одобрили окончательный вариант рукописи.

Авторы заявляют об отсутствии конфликта интересов.

Вклад соавторов: *Каждый автор внес равную долю участия как во все этапы проводимого теоретического исследования, так и при написании разделов данной статьи.*

Linguistic Support of Information Systems and Processes

Igor V. Pertsev, Elena I. Sitnyakovskaya

Siberian State University of Telecommunications and Information Science (SibSUTIS)

Abstract: The article examines the process of dictionary data design, which are sets of lexical units with specific information about them. The methods for creating indexing and search dictionaries are described, which enable efficient organization of search in textual data using specialized indexes.

Keywords: Linguistic support of information systems, design of data dictionaries, indexing and search dictionaries, thesauri and lexical complexes, semantic analysis of textual information, syntactic analysis of textual information, representation of textual information in databases.

For citation: Pertsev I. V., Sitnyakovskaya E. I. Linguistic support of information systems and processes (in Russian). Vestnik SibGUTI, 2024, vol. 18, no. 1, pp. 70-77. https://doi.org/10.55648/1998-6920-2024-18-1-70-77.

Content is available under the license Creative Commons Attribution 4.0 License

© Pertsev I. V. Sitnyakovskaya E. I., 2024

The article was submitted: 15.12.2023;
accepted for publication 10.01.2024.

References

1. Aitchison J., Gilchrist A., Bawden D. *Thesaurus Construction and Use: A Practical Manual*. Aslib-IMI, 2000. 240 p.
2. Shneiderman B., Plaisant C. *Designing the User Interface: Strategies for Effective Human-Computer Interaction (5th Edition)*. Pearson, 2020. 677 p.
3. Beale S. An open-source terminology and knowledge representation system: Ohsumed thesaurus design and implementation. *Computers and Biomedical Research*, 1996, vol. 29, iss. 3, pp. 193-223.
4. Dix A., Finlay J., Abowd G., Beale R. *Human-Computer Interaction (3rd Edition).* Prentice Hall, 2004. 52 p.
5. Hodge G. Systems of thesaurus compilation: a pilot study of thesaurus design and construction. *Journal of Documentation*, 2000, vol. 56(2), pp. 145-168.
6. ISO 25964-1:2011 Thesauri and interoperability with other vocabularies. Part 1: Thesauri for information retrieval.
7. ISO 9241-110:2018 Ergonomics of human-system interaction. Part 110: Interaction principles.
8. Shum S. B., Motta E., Domingue J. Ontology-based knowledge representation for aerospace design. *Journal of Knowledge Engineering and Knowledge Management*, 2001, vol. 19(5), pp. 514-536.

9. Jurafsky D., Martin J. H. *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition (3rd Edition)*. Pearson, 2023. 577 p.
10. Studer R., Benjamins V. R. Knowledge engineering: Principles and methods. *Data & Knowledge Engineering*, 1998. vol. 25(1-2), pp. 161-197.

Igor V. Pertsev

Cand. of Sci. (Engineering), Associate Professor; Associate Professor of the Department of Applied Mathematics and Cybernetics, Siberian State University of Telecommunications and Information Science (SibSUTIS, Russia, 630102, Novosibirsk, Kirov St. 86), phone: +7 383 2698 272, e-mail: pertsev@sibguti.ru, ORCID ID: 0009-0005-2069-3937.

Elena I. Sitnyakovskaya

Cand. of Sci. (Engineering), Associate Professor; Associate Professor of the Department of Applied Mathematics and Cybernetics, Siberian State University of Telecommunications and Information Science (SibSUTIS, Russia, 630102, Novosibirsk, Kirov St. 86), phone: +7 383 2698 272, e-mail: sitlen@sibguti.ru, ORCID ID: 0009-0000-0964-787X.