1. <https://direct.mit.edu/qss/article/2/4/1170/107610/A-meta-analysis-of-semantic-classification-of>
2. <https://arxiv.org/html/2410.09090v1>
3. <https://www.mdpi.com/2076-3417/12/6/3203>
4. <https://direct.mit.edu/dint/article/1/1/58/9974/AMiner-Search-and-Mining-of-Academic-Social>
5. <https://ar5iv.labs.arxiv.org/html/2301.12663>
6. <https://libraryblog.champlain.edu/2024/10/21/semantic-search-research-insights-how-large-language-models-are-changing-library-search-experiences/>
7. <https://www.semanticscholar.org/resources>

**1. Существующие системы семантического поиска**

* **Google Scholar**: Применяет комбинацию методов информационного поиска и анализа цитирований. Для семантической обработки используются нейросетевые алгоритмы NLP (обработка естественного языка) и кластеризация на основе похожести текстов.
* **Microsoft Academic (ранее существовавшая)**: Использовала граф знаний, объединяя информацию о публикациях, цитированиях, авторах и областях исследований. Система могла применять методы обработки знаний и анализа взаимосвязей на основе графов.
* **Semantic Scholar** от Allen Institute for AI: Одна из самых известных систем для семантического поиска. Использует глубокие нейросети, включая трансформеры (например, BERT), и модели для анализа цитирований и взаимосвязей между статьями. Позволяет находить научные публикации с учетом их семантического контекста и цитирования.

**2. Методы и архитектуры, применяемые для реализации**

* **Графы знаний и семантические сети**: Используются для хранения информации о статьях, цитированиях и других метаданных. Графовые базы данных (например, Neo4j) позволяют эффективно связывать и исследовать узлы на основе их семантической близости.
* **Трансформеры и модели на основе BERT**: Современные системы используют модели вроде BERT, RoBERTa и SciBERT (адаптированную для научных текстов) для анализа текстов и семантического поиска. Такие модели позволяют не только находить релевантные статьи, но и учитывать контекст, в котором используются те или иные термины.
* **Сетевые представления и векторизация**: Для семантического анализа применяется векторизация текстов, например, с помощью Word2Vec, GloVe, FastText или моделей Sentence Transformers. Векторизация помогает представлять статьи в виде векторов, что позволяет находить похожие публикации.
* **Анализ цитирований и сеть цитирований**: Важный аспект — учитывать цитируемость и взаимосвязь публикаций. Этот анализ помогает находить статьи, которые были наиболее значимыми для других исследований, выявляя ключевые работы.

**3. Библиотеки для реализации семантического поиска**

* **FAISS от Facebook AI**: Библиотека для поиска по векторным представлениям, часто используется для поиска похожих документов на основе векторов, особенно в случаях, где есть потребность в семантическом сравнении текстов.
* **ElasticSearch с подключением к NLP библиотекам**: ElasticSearch предоставляет возможности для текстового поиска и фильтрации, которые можно расширить за счет моделей машинного обучения. Это решение часто используется для создания поисковых систем с возможностями NLP.
* **LangChain, OpenAI, HuggingFace Transformers**: Платформы и библиотеки, которые предоставляют API и модели для обработки естественного языка и семантического поиска.