**1. Существующие системы семантического поиска**

1. **Google Scholar**

* Характеристики: Использует методы анализа цитирований, семантический анализ текста и ранжирование публикаций.
* Технологии: Применяются нейросетевые алгоритмы NLP, кластеризация текстов и анализ взаимосвязей между публикациями.
* Особенности: Возможность поиска публикаций с учетом их цитируемости и содержания.

1. **Semantic Scholar (Allen Institute for AI)**

* Характеристики: Обеспечивает глубокий анализ текста и цитирований, включая взаимосвязь между статьями и авторами.
* Технологии: Используются модели на базе трансформеров, например, BERT, SciBERT, для семантического анализа научных текстов и граф знаний.
* Особенности: Возможность семантического поиска с учетом контекста и анализа взаимосвязей.

1. **AMiner**

* Характеристики: Фокусируется на анализе академических социальных сетей и взаимосвязей между учеными.
* Технологии: Использует графовые базы данных и алгоритмы анализа сетей цитирования.
* Особенности: Поддерживает исследование межличностных связей в научной среде.

1. **PubMed и ArXiv**

* Характеристики: Применяются в биомедицинских и технических исследованиях.
* Технологии: Используют комбинированные методы, включая машинное обучение и классический поиск по ключевым словам.
* Особенности: Дополняются метаданными, структурированными по тематикам.

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

1. **Графы знаний и семантические сети**

* Описание: Графовые базы данных (например, Neo4j) используются для хранения информации о публикациях, цитированиях и их взаимосвязях.
* Применение:
  + Обеспечивают быстрое и точное нахождение связанных публикаций.
  + Учитывают семантическую близость узлов.

1. **Модели на базе трансформеров**

* Описание: Использование BERT, SciBERT и их модификаций для анализа текста.
* Применение:
  + Семантический анализ текста.
  + Извлечение контекста и взаимосвязей.

1. **Векторизация текста**

* Описание: Преобразование текста в векторные представления с использованием Word2Vec, GloVe, FastText или Sentence Transformers.
* Применение:
  + Поиск похожих текстов.
  + Кластеризация и классификация научных публикаций.

1. **Анализ цитирований**

* Описание: Построение сети цитирования для определения значимости публикаций.
* Применение:
  + Определение ключевых публикаций.
  + Выявление тенденций в научных исследованиях.

**3. Точность существующих систем.**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Система | Precision | Recall | MRR | Особенности |
| Google Scholar | 0.85–0.90 | 0.70 | 0.85 | Хорош для популярных работ, ограничен в семантике. |
| Semantic Scholar | 0.92–0.95 | 0.80 | 0.88 | Высокая релевантность за счет моделей NLP и цитирований. |
| AMiner | 0.88–0.90 | 0.75–0.80 | 0.86 | Сильная сторона — анализ сетей цитирования. |
| PubMed | 0.90–0.95 | 0.85 | 0.89 | Специализация на биомедицине. |
| ArXiv | 0.87–0.90 | 0.80 | 0.85 | Технические и математические публикации. |

**Precision** показывает долю релевантных документов среди всех найденных системой.

**Recall** измеряет долю найденных релевантных документов среди всех существующих релевантных документов.

**MRR** измеряет качество ранжирования результатов.

**4. Проблемы оценки точности**

* Недостаток унифицированных датасетов: Сравнение систем часто затрудняется из-за отсутствия стандартных наборов данных.
* Контекстуальные ограничения: Точность сильно зависит от области поиска, структуры данных и качества запросов.
* Оценка Recall: даже лучшие системы упускают редкие или нецитируемые работы, что особенно важно для новых направлений.

**5. Возможные шаги для повышения точности**

1. **Использование специализированных языковых моделей**

* SciBERT: Адаптирована для научных текстов, учитывает особенности языка статей.
* T5 или GPT с дообучением: Модели генеративного типа могут быть настроены для задач извлечения информации и поиска.

Преимущества:

* Модели, обученные на специализированных наборах данных, лучше справляются с научной терминологией.
* Улучшение Precision за счет семантического анализа терминов и формул.
* Векторизация и улучшение представления данных

1. **Векторизация и улучшение представления данных**

* Sentence Transformers: Модели, такие как SBERT, для нахождения семантически близких предложений.
* Embedding на основе моделей Transformer: Позволяют представлять документы в высокоразмерном пространстве, что улучшает поиск близких по смыслу текстов.
* FAISS (Facebook AI Similarity Search): Ускоряет поиск по векторным представлениям текстов.

Преимущества:

* Улучшение Recall, так как даже менее популярные статьи, схожие семантически, будут найдены.

1. **Графы знаний и цитирования**

* Создание графа, где узлы — это статьи, авторы и ключевые слова, а рёбра отражают связи (цитирование, соавторство, совместные исследования).
* Использование графовых баз данных, таких как Neo4j или TigerGraph, для эффективного хранения и обработки информации.
* Выделение "ключевых" работ с помощью алгоритмов ранжирования (например, PageRank).
* Учет не только прямых цитирований, но и "опосредованных" связей через несколько узлов графа.

Преимущества:

* Повышение Recall, так как учитываются взаимосвязи, не явные при прямом анализе текста.
* Улучшение контекстного понимания цитирования.

1. **Многомодальный поиск**

* Добавление к текстам других источников: изображений, графиков, таблиц и презентаций.
* Применение моделей вроде CLIP (объединяет текст и визуальные данные) для анализа и поиска мультимедийной информации.

Преимущества:

* Увеличение точности поиска по работам, где текстовая информация ограничена

1. **Использование метаданных**

* Включение в поиск информации об авторах, институтах, ключевых словах и источниках.
* Учет метрик цитирования (например, индекса Хирша) для фильтрации и ранжирования.

Преимущества:

* Повышение Precision, так как найденные результаты будут более релевантны интересам пользователя.

1. **Настройка модели**

Дополнительное обучение на специфичных наборах данных

* Использование таких баз данных, как S2ORC (Semantic Scholar Open Research Corpus), PubMed, ArXiv, или сбор собственных данных.

Преимущества:

* Повышение точности при работе с редкими терминами

1. **Анализ ошибок и пользовательских предпочтений**

* Внедрение механизма сбора данных о том, какие результаты были полезны.
* Использование этих данных для обучения моделей ранжирования.

Преимущества:

* Улучшение качества поиска по мере увеличения базы данных предпочтений.