Project Proposal by Thomas Beard
GitHub: https://github.com/Thomas-Beard/IR-Semantic-Analysis-AI
220008104
IN3066

Information Retrieval using Semantic AI Analysis

Problem and Domain

Whilst researching for potential ideas to submit as my project proposal, I came to the realisation that many sources are hard to find especially when there are several trusted search engines (i.e. Google, Google Scholar, Bing, Yandex, DuckDuckGo) that provide different sources, where some are useful, and some are completely irrelevant. To counter this unproductive process, I want to suggest, research, analyse and develop an information retrieval application utilising a form of machine learning that will find the most appropriate resource for a *(A.1)* particular input.

As the rapid growth of online learning platforms has resulted in vast repositories of learning resources, traditional keyword-based search mechanisms often fail to provide relevant results, leading to frustration among students & educators. Many searches return either too many irrelevant results or miss key learning due to the lack of context-aware retrieval. The domain will be based upon education search and e-learning retrieval systems.

And finally, the beneficiaries for this information retrieval application will mostly be used by students and teachers to find relevant sources of information (research papers) to be used in studies.

Test Collection

Given the problem identified, I have come across several sources of data regarding research papers in the computer science field.

Test Collection	**Overview**	**Usefulness**	**Structure**	**Fields & Metadata**	**Topics & Relevance Assessments**
TREC Collections *(C.1)*	Benchmark datasets for IR evaluation in legal, news, medical, and academic search.	Comparing retrieval algorithms and search engines for research papers.	Documents with full-text fields and metadata. Some collections are in XML.	Titles, authors, abstracts, publication sources. Some use Dublin Core metadata.	Includes queries (topics) and relevance judgments.
ArXiv Dataset *(C.2)*	Open-access research papers in AI, ML, and CS. Available in JSON.	Academic search and citation-based ranking.	JSON format with titles, abstracts, authors, full text (for	Structured metadata (DOI, categories, author affiliations).	No built-in relevance labels, but citations and downloads can indicate relevance.

			some papers).		
CORE *(C.3)*	Aggregates open-access research papers worldwide.	Citation-based ranking and academic search.	Papers in PDF, XML, JSON, indexed by repository.	Uses Dublin Core Metadata (titles, abstracts, authors, citations).	No predefined relevance, but citation counts can be used.
Semantic Scholar Open Research Corpus *(C.4)*	Over 39M research papers with metadata and citation graphs.	Search, ranking, citation-based applications.	JSON-based collection with full-text, abstracts, and metadata.	Title, authors, citations, affiliations, structured metadata.	No explicit relevance labels, but citation count, and co-citations indicate relevance.
Microsoft Academic Graph (OpenAlex) *(C.5)*	Large dataset for academic research (formerly MAG).	IR tasks like paper ranking and citation analysis.	JSON, CSV, relational database format.	Titles, abstracts, authors, citations, venue names, DOIs.	No predefined topics, but citations and co-authorship graphs can infer importance.
CiteSeerX *(C.6)*	Citation-based research paper search engine.	Citation analysis and scholarly search.	Indexed papers with metadata, abstracts, some full text.	Uses BibTeX-compatible citation formats.	No topics, but citation counts, and co-citation networks provide relevance.
ACL Anthology *(C.7)*	Computational linguistics and NLP research papers.	NLP and AI-related research retrieval.	Metadata in XML, JSON, or CSV formats.	ACL metadata standards: titles, authors, abstracts, references.	No predefined topics: citations and conference rankings can determine relevance.
S2ORC *(C.8)*	Full-text dataset of academic papers.	IR, citation-based ranking, and	JSON format with full text, citations,	Includes DOIs, citation counts,	No predefined relevance judgments, but citation graphs help.

		NLP search models.	abstracts, references.	author affiliations.	
Kaggle Research Paper Datasets *(C.9)*	Various research paper datasets for IR tasks.	Custom academic search engines, ML-based ranking.	CSV/JSON-based with titles, abstracts, authors, citations.	Metadata varies by dataset.	No predefined topics: user-defined queries can serve as topics.

Choosing a Test Collection

For this project, I have several test collections to choose from and each vary in terms of topics and relevance assessments as some do not contain this structure. My first choice is to focus on using the TREC Collection as the primary test collection. The reason for selecting this dataset is its structured nature, availability of relevance judgments, and alignment with my project's goal of developing an information retrieval system for academic research papers.

Project Proposal by Thomas Beard
GitHub: https://github.com/Thomas-Beard/IR-Semantic-Analysis-AI
220008104
IN3066

References

(A.1) Codex, A.C. (2024). Apache Solr and Machine Learning: Enhancing Search with AI. [online] Reintech.io. Available at: https://reintech.io/blog/apache-solr-machine-learning-enhanced-search [Accessed 18 Feb. 2025].

(B.1) Shah, N. (2018). *ARXIV data from 24,000+ papers*. [online] Kaggle.com. Available at: https://www.kaggle.com/datasets/neelshah18/arxivdataset [Accessed 18 Feb. 2025].

(C.1) trec.nist.gov. (n.d.). *Text REtrieval Conference (TREC) Data*. [online] Available at: https://trec.nist.gov/data.html.

(C.2) kaggle.com. (n.d.). arXiv Dataset. [online] Available at: https://www.kaggle.com/Cornell-University/arxiv.

(C.3) Core.ac.uk. (2025). CORE Services. [online] Available at: https://core.ac.uk/services [Accessed 5 Mar. 2025].

(C.4) Semantic Scholar (2019). Semantic Scholar - an Academic Search Engine for Scientific Articles. [online] Semanticscholar.org. Available at: https://www.semanticscholar.org/.

(C.5) openalex.org. (n.d.). OpenAlex. [online] Available at: https://openalex.org/.

(C.6) Psu.edu. (2016). CiteSeerX Data | CiteSeerX. [online] Available at: https://csxstatic.ist.psu.edu/downloads/data.html [Accessed 5 Mar. 2025].

(C.7) aclanthology.org. (n.d.). ACL Anthology - ACL Anthology. [online] Available at: https://www.aclweb.org/anthology/.

(C.8) allenai (2020). GitHub - allenai/s2orc: S2ORC: The Semantic Scholar Open Research Corpus: https://www.aclweb.org/anthology/2020.acl-main.447/. [online] GitHub. Available at: https://github.com/allenai/s2orc [Accessed 5 Mar. 2025].

(C.9) Kaggle.com. (2025). Find Open Datasets and Machine Learning Projects | Kaggle. [online] Available at: https://www.kaggle.com/datasets?search=research+papers [Accessed 5 Mar. 2025].