Shrinidhi Aarudra

MSDS459 – Knowledge Engineering

Topic Choice and Schema Definition

Consumer Services Sector

Media and Broadcasting

Contents

Abstract

This study outlines the design of a knowledge base and focused web crawler for the Consumer Services sector, particularly in media and broadcasting. The objective is to create a structured, graph-based representation of publicly accessible information on audience behavior, regulations, service models, and broadcast technologies.

The project employs open data sources from organizations like the UK communications regulator and global telecommunications agencies, establishing an initial schema suitable for a graph database. It also details an automated data collection plan using a focused web crawler with relevance filtering, producing a document corpus in JSON Lines format. Ultimately, the knowledge base aims to aid competitive intelligence tasks such as information retrieval, trend analysis, and exploratory question-answering while providing structured inputs for modeling and forecasting.

Introduction

The media and broadcasting services sector is rapidly evolving due to digital distribution, changing audience behaviors, regulatory shifts, and technological advancements. Television broadcasters and digital platforms now navigate a complex ecosystem of linear services, streaming options, and hybrid models. Decision-makers must monitor audience consumption, regulatory policies, and technological trends, but relevant information is often fragmented across various sources.

This research aims to create a knowledge base using a web crawler to consolidate high-quality, publicly available information related to media and broadcasting. By emphasizing open data sources and structuring the information as a knowledge graph, the project facilitates cross-domain queries linking organizations, datasets, technologies, and more.

The knowledge base targets media strategists, broadcast planners, regulatory analysts, and researchers, addressing questions about shifts to streaming, regulatory differences, and technology adoption trends. Ultimately, it is designed for efficient information retrieval, lightweight question-answering, and supporting analytical modeling and forecasting.

Literature Review

Research on focused web crawling provides a framework for targeted information collection in vast environments like the World Wide Web. Chakrabarti, van den Berg, and Dom (1999) introduced focused crawlers designed to acquire relevant documents instead of pursuing exhaustive coverage. They formalized key components, such as topic taxonomy, relevance prediction, and classifier-based filtering, which are essential in domain-specific crawling and competitive intelligence systems. This approach is particularly beneficial for media and broadcasting research, where high-quality information is scattered across various platforms.

Knowledge graphs have emerged as a complementary method for organizing and analyzing web-collected information. They represent entities and relationships that support semantic querying and integration across sources, which is valuable in domains with evolving terminology and complex interdependence like media and broadcasting. By categorizing regulatory bodies, datasets, technologies, and services, knowledge graphs facilitate richer analysis compared to traditional data formats.

Current studies in media analytics often depend on public datasets from regulators, which are typically analyzed in isolation. Integrating these sources into a unified semantic structure enhances discoverability and consistency for longitudinal studies. This project combines focused crawling techniques with a graph-based schema to create a reusable competitive intelligence asset, focusing on applying established methods to a specific domain to help reduce information overload and improve decision-making in the media and broadcasting sector.

Methods

This study uses a design-oriented research approach that combines knowledge graph schema design with a web crawling strategy to improve competitive intelligence in the media and broadcasting services sector. The focus is on defining the domain, structuring data, and collecting data from public web sources, with an emphasis on planning and initial data acquisition rather than full system implementation.

Topic Definition and Rationale

The topic is the Consumer Services sector, focusing on media and broadcasting services, such as television broadcasters and digital media platforms. This area is rich in data, rapidly evolving, and well-documented by open-access datasets. The convergence of audience behavior, policy, and technology makes it ideal for a knowledge graph approach, integrating diverse information sources into a unified structure.

Initial Schema and Knowledge Graph Design

The knowledge base is designed as a graph database for systems like Memgraph, with initial node types including Organization, Document, Dataset, Metric, Service, Technology, Region, and Time Period. Relationships between nodes are explicitly typed and directional, such as "publishes," "describes," and "uses_technology." This schema acts as a lightweight ontology, balancing clarity and practical constraints, while limiting scope to maintain manageability for comparisons and analyses.

Data Sources and Data Collection Plan

Data for the knowledge base will be sourced solely from open web resources, focusing on regulatory research portals, international statistical databases, and technical publications. The project will use a web crawling approach rather than proprietary APIs, with starter URLs defined for each source. Crawling will target research, data, and publication sections for relevance. The crawler will extract text content only, saving the output in JSON Lines files, with each JSON object containing metadata like source URL, publication date, document title, extracted text, and paragraph-level identifiers for future information extraction.

Focused Crawler Design and Relevance Filtering

The focused crawler is designed based on the principles established by Chakrabarti, van den Berg, and Dom (1999). To ensure topic relevance, it employs a combination of pre-download filtering and post-download evaluation. Before downloading, the crawler restricts its traversal to URLs that match specific high-value patterns, such as research reports, datasets, and statistics. Once documents are downloaded, they are assessed using keyword-based relevance heuristics that correspond to a media and broadcasting taxonomy, including topics like audience measurement, streaming services, regulation, and broadcast technology. While a fully trained classifier has not been implemented at this stage, the design anticipates the future integration of supervised relevance classification.

Mapping to the Eight Focused Crawler Activities

Several activities for the focused crawler assignment are in progress. A topic taxonomy has been defined, and high-quality documents from reputable organizations have been identified. Initial topic refinement and web exploration have been conducted manually, alongside curated starter URLs for resource discovery. The distiller and classifier components are conceptually defined but not fully implemented due to time constraints. User evaluation and refinement will be addressed in later project stages.

Intended Users, Applications, and Analytical Use

The intended users of this knowledge base include media strategy teams, broadcast technology planners, regulatory analysts, and academic researchers. Its goal is to provide a

tool for structured querying across various organizations, regions, technologies, and time periods. The knowledge base will facilitate information extraction and lightweight question-answering, while also support future analytical modeling and forecasting with structured inputs from unstructured web documents.

Results

At this stage of the project, our primary results focus on evaluating data sources, determining the feasibility of automated data collection, and assessing the suitability of this information for populating a knowledge graph. An initial survey of publicly available resources confirms that the media and broadcasting services domain is well supported by open-access regulatory and industry data. The identified sources offer consistent insights into audience behavior, service models, regulatory policies, and broadcast technologies, making them suitable foundations for competitive intelligence analysis.

Our exploration of regulatory and industry portals shows significant variability in the structure and granularity of documents and datasets. Regulatory bodies typically publish longitudinal datasets and detailed analytical reports with clear geographic and temporal scopes. In contrast, industry organizations emphasize technical standards, service architectures, and deployment case studies. This diversity highlights the necessity of a graph-based representation, allowing disparate document types, metrics, and concepts to be linked without the constraints of a rigid tabular schema. Notably, most sources provide stable URLs, explicit publication dates, and downloadable content, facilitating reproducible crawling and version control.

Preliminary inspection also indicates that a substantial proportion of relevant documents concentrate within well-defined areas of the web, such as the research and statistics sections of official sites. This supports our focused crawler approach, suggesting that we can construct high-quality document corpora without extensively crawling low-value pages. Furthermore, the recurring presence of common concepts—such as audience measurement, digital distribution, regulatory oversight, and technology transitions—indicates significant overlap across sources, increasing the potential for meaningful entity linking within the knowledge graph. Overall, these results suggest that the selected data sources are sufficiently comprehensive and of high quality to support the construction of an initial knowledge base and to enable subsequent information retrieval, extraction, and analytical tasks.

Conclusion

This study shows that the media and broadcasting services segment of the Consumer Services sector is ideal for creating a focused, graph-based knowledge base for competitive intelligence. High-quality, publicly available datasets from regulatory bodies and industry organizations provide a solid foundation for knowledge extraction. With clear domain boundaries and a constrained schema, this project facilitates the integration of information on audience behavior, service models, regulations, and broadcast technology

Preliminary results confirm that a focused crawling strategy is effective, as relevant documents tend to be located in specific areas of the web, minimizing information overload. However, differing terminology and metric definitions across sources underscore the need for a knowledge graph approach to manage semantic diversity without premature standardization. These insights address the challenge of organizing fragmented information into a coherent, quarriable structure for monitoring and comparison over time and across regions.

Challenges remain, including reliance on heuristic relevance filtering in the crawler, which can lead to the inclusion of less relevant documents, and the need for ongoing schema refinement as modern technologies emerge. Despite these limitations, the proposed knowledge base is poised to enhance information retrieval, extraction, and exploratory questioning, supporting downstream analytical modeling and forecasting. Overall, this project lays a strong groundwork for a scalable competitive intelligence system utilizing open data and targeted web mining techniques.