Cross Media Entity and Concept Driven Search

Chalitha Perera
Zaizi
cperera@zaizi.com

Dileepa Jayakody
Zaizi
djayakody@zaizi.com

ABSTRACT

In recent years there is a rapid growth of unstructured text and multimedia content, which includes audio, video and image content on the web and within the enterprise. Handling these large volumes of data is cumbersome without effective methods for content analysis and retrieval. This paper presents Sensefy – a cross media information retrieval system, which uses higher-level semantic concepts and entities for retrieving documents of different media types. This paper describes the process of automated semantic information extraction from multimedia documents, which is central in developing concept and entity driven search across different media types. Further, the Media In Context (MICO) platform is described as a platform for cross media analysis.

CCS Concepts

• Information systems→ Information Retrieval→ Specialized information retrieval→ Multimedia and multimodal retrieval

Keywords

Multimedia Information Retrieval, Entity Search, Semantic Search

1. INTRODUCTION

Knowledge and information are the driving forces behind the success of any industry leading enterprise today. Staggering amounts of content get accumulated in enterprises on a daily basis. Most of this content is in the form of unstructured texts such as emails, presentations and business documents, which are written in natural language narrative and multimedia files. This content doesn't conform to any predefined data model or structure but contains knowledge vital to the enterprise. Unveiling the hidden knowledge in this unstructured content has become a huge challenge in the enterprise today. Searching relevant documents with accuracy and efficiency is another major challenge. For successful business operations, it is essential that a knowledge worker can quickly locate relevant content and have access to the required knowledge to perform their daily activities. Therefore, information and knowledge discovery from the organization's content is now a key success factor for any knowledge driven enterprise.

Over the years, search solutions have evolved to cater to growing user expectations and enterprise search needs. Today, enterprise search solutions provide federated search across all content repositories in the enterprise, preserving the user access permissions of the heterogeneous content repositories. With the emergence of linked-data, semantic web and machine learning technologies, enterprise search has evolved beyond simple keyword search to intelligent semantic search. Today, the users of information retrieval systems can search and navigate through unstructured content using concepts and entities that are contextually relevant to their query. To enable such advanced semantic search, the content needs to be pre-processed and analyzed to extract hidden semantic knowledge. For text-based content, this is achieved using Natural Language Processing (NLP) to detect nouns and other keywords that represent real world concepts and entities, and link them to contextual knowledge maintained in knowledge bases (E.g. DBpedia, Freebase, YAGO or enterprise specific knowledge bases). This process enhances unstructured text content by adding semantic metadata and constructing knowledge graphs with linked entities, providing contextual meaning to the documents. This additional semantic metadata is instrumental to provide intelligent entity driven and concept driven search capabilities. STICS [1] is one such semantic search system, which allows users to search online news articles by using entities.

STICS semantic search is limited to text documents. In the enterprise, documents include audio, video and image files in addition to text documents. Also enterprise documents such as presentations and web pages contain embedded multimedia files. Although semantic analysis and search with textual content has advanced significantly, knowledge extraction and semantic search for multimedia content is still at research level. This is mainly due to the lack of compatibility, analysis methods for extracting knowledge from multimedia, and the lack of standard semantic annotation models for annotating multimedia with semantic metadata.

Sensefy is an open-source enterprise search application that provides cross media entity and concept driven search with the use of the Media In Context (MICO) platform [2]. MICO provides an extensible framework for cross-media analysis, meta-data publishing and multimedia querying to facilitate semantic enhancements to textual, image, audio and visual content. By combining MICO as the semantic analysis framework, Sensefy aims to provide entity and concept driven search features with multimedia content in an intuitive user interface.

The rest of the paper is organized as follows: Section 2 gives an overview of the Sensefy architecture. It describes the main components of the architecture and their functions in sub sections. Section 3 gives the conclusions and the future work planned for the project.

SEMANTICS 2016: Posters and Demos Track
September 13-14, 2016, Leipzig, Germany

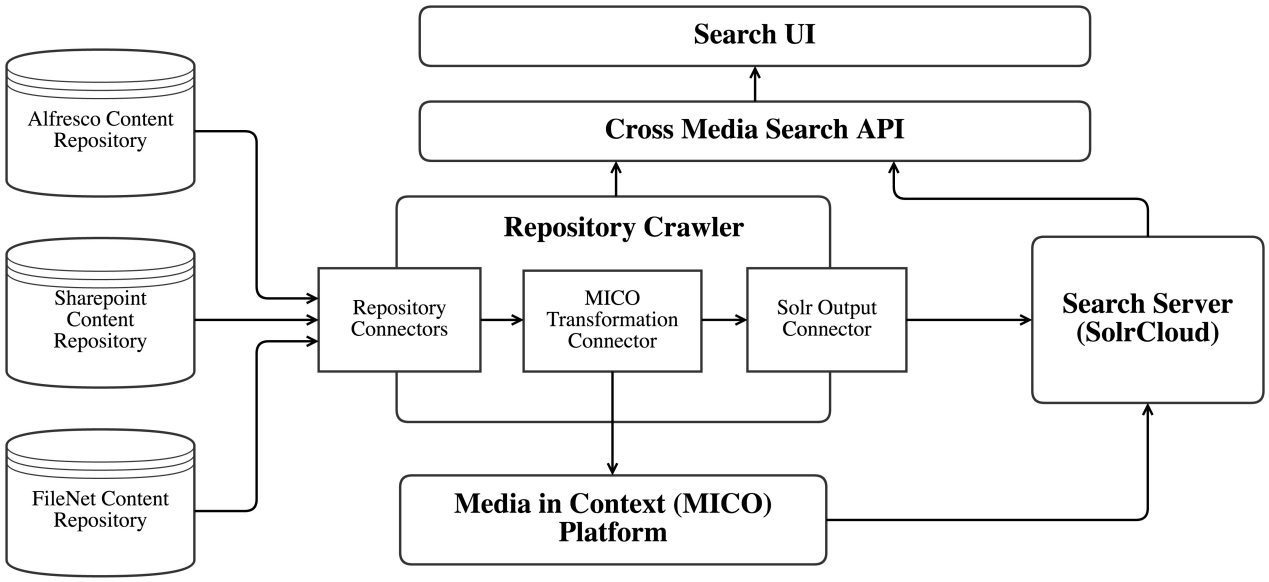

Figure 1. Schematic Diagram of Sensefy Architecture

2. SENSEFY ARCHITECTURE

The system architecture for Sensefy aims to address the following set of requirements.

- *Semantically enhance unstructured document content* – Provide structure to unstructured content by extracting semantic features including named entities.
- *Search with real world concepts and entities* – Extend traditional keyword based search and provide the ability to search documents using concepts and entities such as people, organizations and locations.
- *Search across different media types* – Extend entity and concept driven search to find videos and images in addition to text content.
- *Provide smart search assistance* - Provide user friendly search features such as autocomplete, suggestions and highlights with entities and concepts
- *Integrate different content repositories to provide federated search solution* – Users should be able to search content in different content repositories using a single search query.

Figure 1 shows the key components of the Sensefy architecture and the following sections will describe each component in detail.

2.1 Media In Context (MICO) Platform

MICO is a platform for performing the cross media extraction, analysis, metadata publishing and querying. MICO provides a set of extractors that can be arranged in a pipeline to extract required

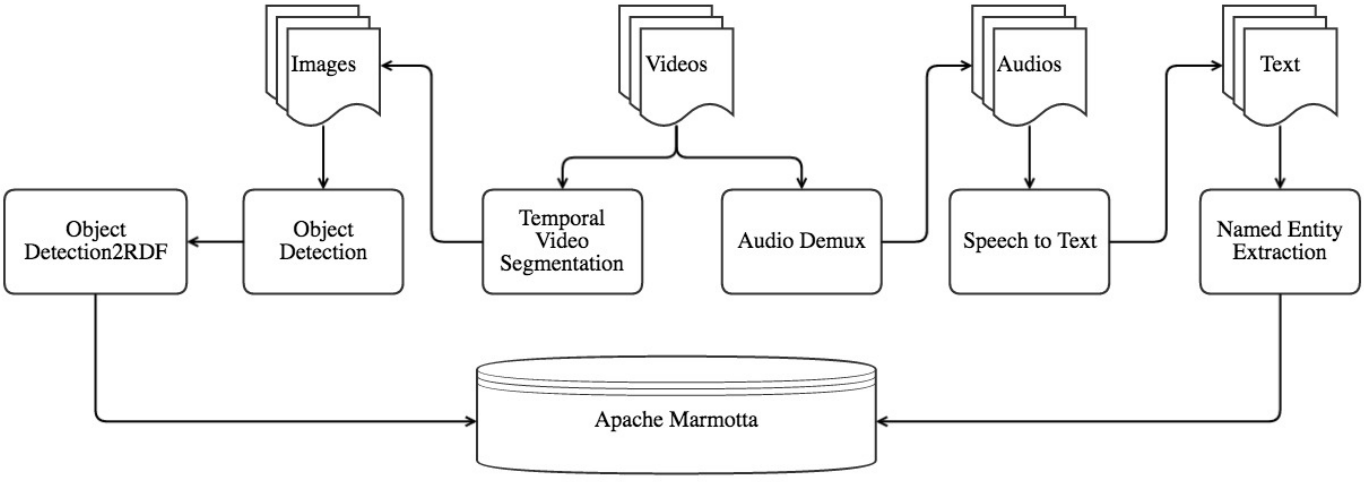

Figure 2. Extractor Pipeline Used in Cross Media Extraction Process

metadata from multimedia files. Extracted metadata is stored as RDF triples in Apache Marmotta backend. Metadata can be queried using SPARQL or LDPath queries. MICO provides a java library named "Anno4j" [3], which provides a high-level API to programmatically create and query multimedia metadata using Web Annotation Data Model [4]. Anno4j lets the developers access the metadata model and perform required cross media analysis avoiding the need to write complex SPARQL queries to handle the linked-data.

The extractor pipeline used for cross media extraction in the Sensefy use case is given in Figure 2.

- Temporal Video Segmentation – Segments a video file into set of key frame images.
- Audio Demux – Extracts audio stream from a video and acts as a preprocessing step for speech to text extractor by changing the sampling rate of the audio.
- Speech to Text – Converts the audio stream into text. This extractor is supported by a Diarization extractor, which acts as a preprocessor by splitting large audio files into more manageable ones.
- Named Entity Extraction – Performs named entity recognition and linking based on Stanford NLP and OpenNLP language models.
- Object Detection – Recognizes objects in images. Object detector provides a generic object recognition extractor that can be trained for different domains with training image dataset and once trained it is capable of identifying objects in a given image.

For more details about the MICO platform and the data-model, readers are encouraged to refer MICO Publication Volume 4 [5].

2.1.1 Named Entity Extraction and Linking

Named entity extraction is one of the most important concepts in information extraction and natural language processing. A named entity is anything that can be identified as a proper noun. In general, the task of named entity recognition focuses on identifying proper nouns and classifying them into several categories.

In MICO, we extract people (e.g. Barack Obama), locations (e.g. Sydney), organizations (e.g. Real Madrid C.F.) and concepts (e.g. soldier) from unstructured text, including text streams extracted from audio and video files.

Entity linking is the process of linking identified named entities (mentions) with real world entities, which reside in knowledge graphs. Each entity in a knowledge base has a unique URI (Uniform Resource Identifier). With entity linking it is possible to enrich document content with all the rich metadata related to entities and concepts in knowledge bases. Using MICO we can semantically enhance documents by linking with publicly available knowledge bases such as DBpedia, YAGO and also have the capability to plug-in more domain specific custom built knowledge bases for semantic enrichment.

Entity linking also provides a solution to the ambiguity problem in entities. Given a mention for example "Queen" it is important to know which real world entity this mention refers to. Is it "Queen Elizabeth", "Queen the band", "Queen Cafe" or any other Queen? It is possible to solve the disambiguation problem by considering the context in which the mention appears alongside with information extracted from knowledge graph. Entity disambiguation plays a central part in improving the precision of entity driven search results.

2.2 Repository Crawler Framework

For the purpose of crawling heterogeneous content repositories we use the Apache Manifold Connectors Framework (ManifoldCF) [6], which is a flexible framework for connecting different source content repositories with different target indexes. ManifoldCF defines a security model that enables enforcing security policies of source repositories. It also provides a scheduling service for

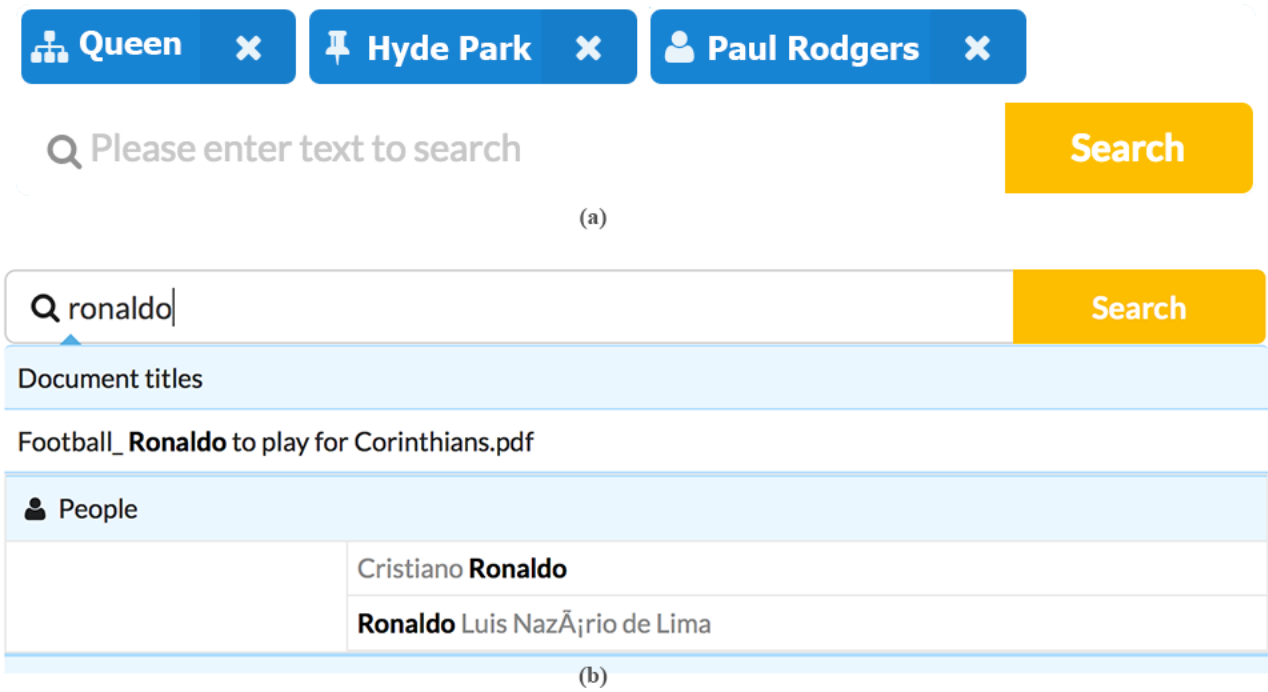

Figure 3. (a) Searching with combining entities in Sensefy UI (b) Disambiguated entity suggestions

running crawler jobs. There are four main types of connectors in ManifoldCF.

- *Repository Connectors* - Allows you to connect to content repositories for document crawling purposes.
- *Authority Connectors* - Provide authority services such as retrieving access tokens for a given user enforcing security policies of content repositories.
- *Transformation Connectors* - Allows you to modify content fetched from repositories before indexing into search servers.
- *Output Connectors* - Allows you to index documents, metadata and permissions in different search servers (indexes) such as Apache Solr, ElasticSearch and Amazon CloudSearch.

We have implemented a transformation connector in the ManifoldCF framework to enhance multimedia documents by sending them to the MICO platform to extract metadata. Metadata querying from the platform is done leveraging Anno4j.

2.3 Sensefy Search API and UI

Sensefy Search API primarily handles the cross media search queries with the search server. In Sensefy, the search server used at the core is Apache Solr. It is deployed as a Solr Cloud for high availability and high performance.

The search API provides smart search services such as search autocomplete and suggestions for keywords, entities and concepts to allow users to search documents with contextual insight. The Sensefy Search user interface (UI) presents the results retrieved, in a unified and intuitive manner. Search UI provides smart suggestions on the go, enabling the user to select a combination of entities and concepts to build a complex search query with ease (Figure 3(a)). Using the smart filters, a user can filter the retrieved results based on different attributes of the content such as media type, file size range, source repository and language.

One of the key benefits of using entity and concept driven search with smart suggestions is that it can produce results with improved precision compared to traditional keyword search. For example, search results from a keyword based query for search term "ronaldo" will contain document related to "Cristiano Ronaldo" the Portuguese player as well as the documents mentioning "Ronaldo" the Brazilian player. Here the user's intent might be to search for either one of them but not both. Using the entity suggestions feature, the user can select relevant entities from the list of suggestions, to limit the results that would contain only the documents mentioning the selected entities (Figure 3(b)).

3. CONCLUSION AND FUTURE WORK

In this paper we presented Sensefy, an enterprise cross media search solution, which allows you to search different media content with real world entities and concepts. We have explained the components of the Sensefy architecture and also have described MICO as a platform for cross media analysis.

We have explained how the different extractors in MICO platform can be configured in the analysis pipeline to extract required metadata from content, which is central in powering the cross media search solution of Sensefy. Extracted metadata are indexed with the documents enabling the entity driven search in Sensefy. We have shown that it is possible to get search results with higher precision by using Sensefy's entity driven search combined with smart suggestions and entity disambiguation. To our knowledge, Sensefy is one of the first systems to provide semantic concepts and entity based search across different media types.

Current implementation of entity and concept driven search is limited to videos, audios and text documents and does not provide great support for images. Current object extractor in MICO is limited to identifying animals in a given image. Therefore one of the major future additions to Sensefy will be the automatic semantic concept (object) extraction and annotation from images.

4. ACKNOWLEDGMENTS

MICO (Media in Context) is a research project partially funded by the European Commission 7th Framework Program (grant agreement no: 610480). This project integrating MICO platform with Sensefy is carried out as part of the exploitation of MICO framework. We would like to acknowledge all the partners of the MICO project for their continuous support in making this project a success. Further we like to acknowledge the members and contributors of the open source projects, Apache ManifoldCF, Apache Solr for their contributions.

5. REFERENCES

[1] Hoffart, Johannes, Dragan Milchevski, and Gerhard Weikum. *STICS: searching with strings, things, and cats* in Proceedings of the 37th international ACM SIGIR conference on Research & development in information retrieval. ACM, 2014.

[2] Aichroth, P., Weigel, C., Kurz, T., Stadler, H., Drewes, F., Björklund, J., Schlegel, K., Berndl, E., Perez, A., Bowyer, A., and Volpini, A. *MICO - Media in Context* in Proceedings of the IEEE International Conference on Multimedia & Expo Workshops (ICMEW), 2015.

[3] Anno4j/anno4j. 2016. GitHub. https://github.com/anno4j/anno4j

[4] Web Annotation Data Model. 2016. World Wide Web Consortium https://www.w3.org/TR/2016/CR-annotation-model-20160705/

[5] Aichroth, P., Björklund, J., Schlegel, K., Kurz, T., and Köllmer, T. *Specifications and Models for Cross-Media Extraction, Metadata Publishing, Querying and Recommendations:Version II*, Salzburg Research Forschungsgesellschaft mbH, Salzburg, Austria, 2015.

[6] Welcome to the Apache ManifoldCF™ project!, 2016. Manifoldcf.apache.org. https://manifoldcf.apache.org