DOCTORAL ABSTRACT

Search and Analytics Using Semantic Annotations

Dhruv Gupta
Max Planck Institute for Informatics, Saarbrücken, Germany
dhgupta@mpi-inf.mpg.de

Abstract

Current information retrieval systems are limited to text in documents for helping users with their information needs. With the progress in the field of natural language processing, there now exists the possibility of enriching large document collections with accurate semantic annotations. Annotations in the form of part-of-speech tags, temporal expressions, numerical values, geographic locations, and other named entities can help us look at terms in text with additional semantics. This doctoral dissertation presents methods for search and analysis of large semantically annotated document collections. Concretely, we make contributions along three broad directions: indexing, querying, and mining of large semantically annotated document collections.

Indexing Annotated Document Collections. Knowledge-centric tasks such as information extraction, question answering, and relationship extraction require a user to retrieve text regions within documents that detail relationships between entities. Current search systems are ill-equipped to handle such tasks, as they can only provide phrase querying with Boolean operators. To enable knowledge acquisition at scale, we propose GYANI, an indexing infrastructure for knowledge-centric tasks. GYANI enables search for structured query patterns by allowing regular expression operators to be expressed between word sequences and semantic annotations. To implement GREP-like search capabilities over large annotated document collections, we present a data model and index design choices involving word sequences, annotations, and their combinations. We show that by using our proposed indexing infrastructure we bring about drastic speedups in crucial knowledge-centric tasks: 95× in information extraction, 53× in question answering, and 12× in relationship extraction.

Hyper-phrase queries are multi-phrase set queries that naturally arise when attempting to spot knowledge graph facts or subgraphs in large document collections. An example hyper-phrase query for the fact ⟨MAHATMA GANDHI, NOMINATED FOR, NOBEL PEACE PRIZE⟩ is: ⟨{*mahatma gandhi*, *m k gandhi*, *gandhi*}, {*nominated*, *nominee*, *nomination received*}, {*nobel peace prize*, *nobel prize for peace*, *nobel prize in peace*}⟩. Efficient execution of hyper-phrase queries is of essence when attempting to verify and validate claims concerning named entities or emerging named entities. To do so, it is required that the fact concerning the entity can be contextualized in text. To acquire text regions given a hyper-phrase query, we propose a retrieval framework using combinations of n-gram and skip-gram indexes. Concretely, we model the combinatorial space of the phrases in the hyper-phrase query to be retrieved using vertical and horizontal operators and propose a dynamic programming approach for optimized query processing. We show that using our proposed optimizations we can retrieve sentences in support of knowledge graph facts and subgraphs from large document collections within seconds.

Querying Annotated Document Collections. Users often struggle to convey their information needs in short keyword queries. This often results in a series of query reformulations, in an attempt to find relevant documents. To assist users navigate large document collections and lead them to their information needs with ease, we propose methods that leverage semantic annotations. As a first step, we focus on temporal information needs. Specifically, we leverage temporal expressions in large document collections to serve time-sensitive queries better. Time-sensitive queries, e.g., *summer olympics* implicitly carry a temporal dimension for document retrieval. To help users explore longitudinal document collections, we propose a method that generates time intervals of interest as query reformulations. For instance, for the query *world war*, time intervals of interest are: $[1914, 1918]$ and $[1939, 1945]$. The generated time intervals are immediately useful in search-related tasks such as temporal query classification and temporal diversification of documents.

As a second and final step, we focus on helping the user in navigating large document collections by generating semantic aspects. The aspects are generated using semantic annotations in the form of temporal expressions, geographic locations, and other named entities. Concretely, we propose the xFactor algorithm that generates semantic aspects in two steps. In the first step, xFactor computes the salience of annotations in models informed of their semantics. Thus, the temporal expressions *1930s* and *1939* are considered similar as well as entities such as usain bolt and justin gatlin are considered related when computing their salience. Second, the xFactor algorithm computes the co-occurrence salience of annotations belonging to different types by using an efficient partitioning procedure. For instance, the aspect $\langle$\{usain bolt\}, \{beijing, london\}, $[2008, 2012]\rangle$ signifies that the entity, locations, and the time interval are observed frequently in isolation as well as together in the documents retrieved for the query *olympic medalists*.

Mining Annotated Document Collections. Large annotated document collections are a treasure trove of historical information concerning events and entities. In this regard, we first present EventMiner, a clustering algorithm, that mines events for keyword queries by using annotations in the form of temporal expressions, geographic locations, and other disambiguated named entities present in a pseudo-relevant set of documents. EventMiner aggregates the annotation evidences by mathematically modeling their semantics. Temporal expressions are modeled in an uncertainty and proximity-aware time model. Geographic locations are modeled as minimum bounding rectangles over their geographic co-ordinates. Other disambiguated named entities are modeled as a set of links corresponding to their Wikipedia articles. For a set of history-oriented queries concerning entities and events, we show that our approach can truly identify event clusters when compared to approaches that disregard annotation semantics.

Second and finally, we present jigsaw, an end-to-end query-driven system that generates structured tables for user-defined schema from unstructured text. To define the table schema, we describe query operators that help perform structured search on annotated text and fill in table cell values. To resolve table cell values whose values can not be retrieved, we describe methods for inferring null values using local context. jigsaw further relies on semantic models for text and numbers to link together near-duplicate rows. This way, jigsaw is able to piece together paraphrased, partial, and redundant text regions retrieved in response to structured queries to generate high-quality tables within seconds.

This doctoral dissertation was supervised by Klaus Berberich at the Max Planck Institute for Informatics and htw saar in Saarbrücken, Germany. This thesis is available online at: `https://people.mpi-inf.mpg.de/~dhgupta/pub/dhruv-thesis.pdf`.