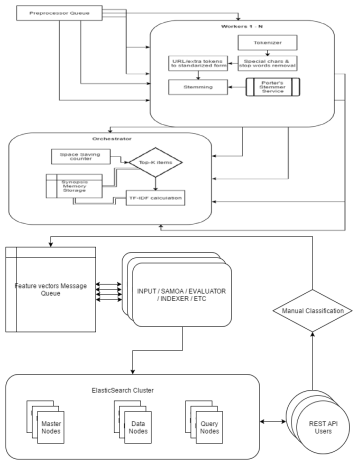
Social Media Spam & Content Quality

Research Student
Ilias Bertsimas

Supervisors
George Magoulas
Dell Zhang



Department of Computer Science and Information Systems

Research Aims

Our research is motivated by the problems arising from dealing with social media data and their streaming data nature. Each source has no standard schema or no schema at all. Also, traditional machine learning is largely batch based and not geared towards real-world data streams. As a result, such machine learning is not real time in a real world scenario making the usefulness of the data obsolete. To address these challenges, we are developing techniques for **schema consolidation in standardized format** and a **machine learning pipeline reference architecture for data streams.**

Research Methodology

We undertake schema remapping of heterogeneous social media data sources through the use of **semantic similarities** between them, which we translate to a **Friend-Of-A-Friend (FOAF) ontology** format, using a schema mapping reference for each of the social media sources (see Figure 1). This is part of the pipeline of the reference architecture we have developed where a continuous flow of data from social media data sources all the way to the machine learning result exists and can provide a near-real-time machine learning processing over the data stream.

Research Approach

We evaluated a lot of architectures from typical batch processing machine learning solutions to streaming machine learning and reached the conclusion that a concrete real-world pipeline reference architecture is lacking. We designed such a pipeline reference architecture around the **apache SAMOA** project. We reap the potential benefits of apache SAMOA and its abstraction of popular machine learning algorithms over multiple streaming processing engines. We created a framework that contains a modified apache SAMOA that provides the necessary components for a scalable highly performing streaming machine learning pipeline (see Figure 2). This will allow us to plug into real-world social media streams and provide near-real-time results. We aim to complete our pipeline reference architecture implementation details and test using real world streams to better evaluate the performance of our machine learning solution along with the ability to successfully handle real-world scenarios.

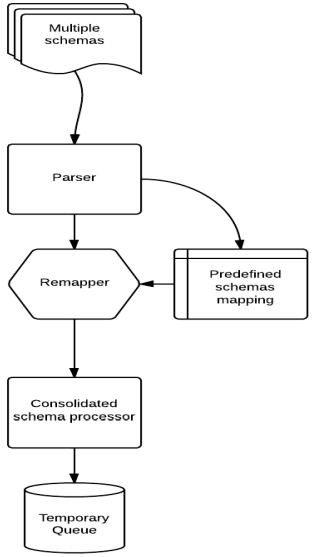


Figure 1. Schema Consolidation Process

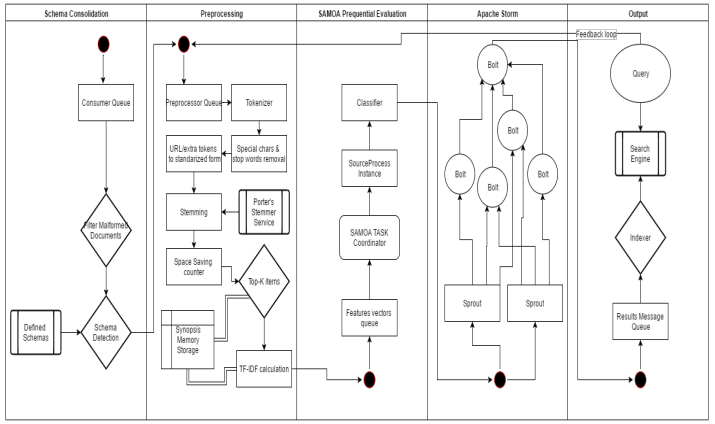


Figure 2. Reference Pipeline Architecture

Publications

Ilias Bertsimas, George Magoulas, A near-real-time machine learning pipeline architecture for streaming social media data (pending submission)