

Abstract

From user point of view, TIMESWEN is a behavior based web portal designed to provide popular articles and news stories at real time. TIMESWEN serves second fastest pure trending content to readers. Important things from system are various state of the art Data storage technologies with Data Relation mapper, Pattern learning algorithms, Rank & indexing algorithms and Natural Language processor.

System flow starts with crawling batch of URLs from famous high ranked news providers in repetitive manner. Once data has been offline on server, it fires a crone job which includes a sequence of controllers and asynchronous script-bots. Every controller has been made to apply its own specific series of algorithms. Parser-bot applies Regular Expression Match on DOM structure of crawled data for predicting the relevant content. It also learns and becomes more and more accurate form past experience to find relevant textual data. Relation mapper controller then applies NLP on parsed content to map duplicity of same news or similar article, and removes it based on its popularity score. It also binds relation between tokens using N-gram Model and saves it into a graph based data storage technology called Neo4J. Because of the Neo4J is connected with Google's knowledge graph, it becomes fast and easy to map words with N-gram model. Algorithm similar to Edge-rank indexes news & stories according to user's place, past searches and behavior.

Finally, web server uses this indexed data to render news and articles by implementing one of the best UX guidelines provided by Google Inc. known as The Material Design.