Ideation Phase - Literature Survey

Project Domain	Cloud Application Development
Project Title	News Tracker Application
Team ID	PNT2022TMID40746
Date	3rd Sept 2022

Author	Martijn Kleppe and Marco Otte
Title	Analysing and understanding news consumption patterns by tracking online user behaviour with a multimodal research design
Year	2017

Understanding users' online behaviour is of growing interest to academic researchers in a variety of fields. The Newstracker is a custom-built system that collects Web activities of specified and authenticated users, cleans the data by removing nonrelevant data, extracts the associated content, and stores this as a new data set to be used for analysis. To create the Newstracker, we applied two out of several data mining techniques: Web usage mining and Web content mining. While designing the Newstracker, we encountered several issues particularly during the pre-processing phase. First, caching is cited as an issue for reliably collecting user Internet use. When using a cached Web page, a browser will not request the page or elements of the page and thus not produce the HTTP request needed for the proxy server to log activity. The basis of the Newstracker was a local proxy. Even though applying a proxy to monitor Web behaviour is a relatively easy and a low-cost solution, it does ask some technical knowledge of the respondent. A second methodological challenge is setting up the whitelist of websites that need to be monitored. A third and possibly most important methodological challenge is to find respondents who are willing to have their online behaviour being monitored. In the data analyses phase, we expect the future analyses of the scraped content will be a challenge. Especially researchers who are monitoring the behaviour of large amounts of respondents during longer period will gather large amounts of scraped content.

Author	Swit Phuvipadawat, Tsuyoshi Murata
Title	Breaking News Detection and Tracking in Twitter
Year	2010

Twitter is a social networking service that allows users to share information, which is described by Twitter as "What's happening?" in a form of short texts (140 characters). Keywords-Twitter, Topic Detection and Tracking, Real-time text-mining, Information Retrieval. Breaking news is defined by Wiktionary as "news that has either just happened or is currently happening. Breaking news may contain incomplete information, factual error or poor editing because of rush." With this definition Twitter can fit the needs of breaking news delivery. However, news posted in Twitter requires an effort to discover it. Firstly, users often have problems of deciding which users to follow. That is, to find users with interesting tweets. Secondly, users need to read through status updates and follow links to obtain further information. There are two important elements in a message: emotions and facts. The inclusion of emotions in the message makes news delivered in Twitter, different from news delivered by professional journalists. In this paper, we present a method to collect, group, rank and track breaking news. Tasks are divided into two stages: story finding and story development. In this paper, we focus on facts part of messages. Emotions are left for our future works. Tasks are presented in three steps: sampling, indexing and grouping. We use the Stanford Named Entity Recognizer (NER) for the classification of proper nouns. NER provides a general implementation of linear chain Conditional Random Field (CRF) sequence models, coupled with well-engineered feature extractors for Named Entity Recognition. In this work, we discussed the characteristics of breaking news in Twitter and presented a method to collect, group, rank and track breaking news from Twitter. To improve the similarity comparison for short-length messages, we put an emphasis on proper nouns. There are several factors to rank the news story. In this experiment we use reliability, popularity and behave for the ranking factors. With these counterparts, we can understand the user's perception to news stories, impacts to the mass audience and the pattern in which the information spreads.

Author	Milos Krstajic´, Mohammad Najm-Araghi, Florian Mansmann and Daniel A Keim
Title	Story Tracker: Incremental visual text analytics of news story development
Year	2013

Understanding temporal development of unstructured and semi-structured text data streams is becoming increasingly important in many application areas, such as journalism, politics, or business intelligence. Keywords are News stream analysis, topic evolution, dynamic visualization, text analytics. Research on topic development, visualization, and analysis of temporal dynamics in information streams has strong connections to the fields of text mining, information visualization, and time-series analysis. A well-known initiative in text mining was topic detection and tracking (TDT),1 which investigated methods for discovering events in broadcast news streams. The news articles are collected as XML data, as described by Krstajic´ et al. Each time-stamped data item contains metadata, such as named entities or tags. Entities identify people or organizations that are mentioned in the document, while tags categorize the news (e.g., earthquake, sports). Very often, news stories have braided nature, that is, documents that belong to different clusters can be still highly related. Furthermore, news stories may split into two different topics at some point. The visualization in our system can be seen as like parallel coordinates, which is a well-known technique for visualizing highdimensional data. The global overview visualization helps in identifying the major news stories over a long period of time, and it is enriched with the interaction techniques to filter and rerank stories on various user-adjustable criteria in order to provide a clutter-free display.