Impactfulness

According to the FICCI-EY report, the media and entertainment business is estimated to grow 25% to reach Rs. 1.73 trillion (US$ 23.29 billion) in 2022. Millions of people search for the news/articles/papers which are trending or viral on Google or similar search engines. But search engines start to repeat search results after the third page. This causes inflated importance of some posts (going “viral” unnecessarily) and gives a noisy Google search experience that may be hiding more relevant news articles from end-users. Google in its policies clearly mentions (https://support.google.com/news/publisher-center/answer/9607104) to not artificially freshen stories. It states that : If an article has been substantially changed, it can make sense to give it a fresh date and time. However, it’s against our guidelines to artificially freshen a story when the publisher didn't add significant information or demonstrated a compelling reason. The reason being the amount of data the we have to deal with everyday. There it becomes important to have an efficient solution that reduces noise search throughout the internet medium so that users can get updated and accurate information that they can process and make use of.

Innovativeness

We used Graph and Hyper-node graph to solve the problem. Graph helps us to traverse through the existing articles and their similarities quickly as compared to traditional database. The Hyper-node indicates the metadata for a cluster of duplicate or near-duplicate articles/posts, and how their metadata relates to each other. The individual articles and their metadata would be clustered together as relations to the hyper-node. As a result, each hyper-node represents all duplicate versions of the articles and posts, normalizing their representation. This hyper-node and its metadata can then be used to group articles/posts together in a search application to minimize noisy search for news articles/posts and help end-users identify if an article/post is actually “going viral” or just overhyped and not worth their time.

Hypergraph is a less travelled road in graph data processing and could be of significant use in future where we have not just news but large documents and textual material to deal with. Clustering with hypergraphs is a domain that can be explored in graph databases.

Instead of using the already available graph query machine learning algorithms we tried to implement the BERT (Bidirectional Encoder Representations from Transformers) by Google AI for language modelling. BERT’s key technical innovation is applying the bidirectional training of

Transformer, a popular attention model, to language modelling. This is in contrast to previous efforts which looked at a text sequence either from left to right or combined left-to-right and right-to-left training. The detailed workings of Transformer are described in a paper by Google.

Ambitiousness

The main advantage of using graphs over traditional database system is the query performance. As the query performance in graph doesn't dependent on amount of data but one the number of concrete relationship. In our graph, there is only one type of vertex, that is "Article" which contains the metadata of article. The edges in the graph represent similarity between different vertices(Articles). There are total 97 vertices in the beginning, and a new vertex is added every time the user uses the application. Graph Neural Network is an ambitious approach to solve real world problems. Our schema is a small representation of the challenge that we are catering to.

We have around 97 pieces of information drawn into one article vertex and another vertex for its publisher. On top of that we have a hypernode structure that clusters all similar data together along with keeping its important information intact in form of metadata. We utilise the normalized metadata for further similarity computing between hypernodes.

Applicability

Different organisations ranging from small startups to big MNCs can use our project for various purposes like duplicate document identifier, copyright detection, retrieving similar documents/articles. Industries relating to search engines, News publisher, Research paper publisher are our main targets. Organisations like Google, Arxiv, IEEE can use our project to filter out duplicate articles. Another use would be to check how similar the new article/paper is compared to old content and also list the old content that they are similar to if any. Graphs can easily be extended to any language modelling tasks that human can think of.