CS-410 Text Information System – technical review

Analysis of different Information Text retrieval methods in text mining

Introduction:

Data is growing enormously due to transformations such as handling everything electronically, it reaches all over the world even in a very small village, it is unpredictable to judge the volume of the data in the rapidly growing fashion. Data can be stored in database or text form. When it comes to a database , it is stored in a structured way and it can be retrieved easily by using a unique query. In the text form, the data can be stored in neither unstructured nor Semi-structured form. I am always surprised at how the data is getting retrieved when it is stored in unstructured form. how the existing technique is well around handling huge volume with retrieving the right data. I start to do some analysis on the existing technique and find out how it works better with a huge volume of data.

Body Paragraphs:

The main goal of the information retrieval should be retrieving the correct information which is more relevant to the user’s query. There are multiple techniques that can be used to retrieve the document. Here is the same technique mentioned below for the analysis.

A vector space model is based on a similarity notion model, initially, we represent the text documents as vectors then we convert them into a numerical format. It mainly used the bag of words concept. Before processing, it removes the stop words, special characters, punctuation etc. It defines the term document matrix and term weights on each document. Term-document matrix is created based on the presence of the query word, then term weights are assigned to the matrix for all the documents. Sometimes the words which present most of the documents might not contribute to the relevance, but less occurring words across the document can decide the relevance. It can be achieved using the term frequency-inverse document frequency.

Data can be retrieved based on the user’s query. It is being measured using two important terms:

Precision: the percentage of the retrieved document which is expected to be relevant to the User’s query

Relevant: The percentage of the retrieved document which is relevant to the user’s query. The two basic measures of accuracy are precision and recall.

It is not possible to have perfect precision and perfect recall in all the scenarios, it compromises each other to achieve higher precision we may have to lose the recall or vice versa. We cannot consider the single point of precision or recall achieving the ranking. Better to plot the precision-recall curve to decide the right approach. The assumption is always some sacrifice between precision and recall.

Natural language processing typically involved with large set rules is coded. Modern NLP is mostly based on machine learning.it is mostly based on algorithms.

The standard measure means average precision (MAP) is described as the average precision at all points which bring the relevant documents. Here the recall value will be high when the precision is lower than zero. NDGC- Normalized Discounted Cumulative Gain: It considers multi-level relevance. First, it considers the gain as the correspondence to the level of relevance. It measures the overall utility of the n documents by the sum of the gain of the corresponding relevant document. It discounts the gain of each document that is ranked low, so highly ranked documents can be counted towards gain. It uses the ranking to find the upper bound and then normalize the actual gain with this calculated upper bound. NDCG is a better approach compared to MAP. But it is hard to judge the discounting strategy, it can become like MAP in some cases. MAP can play a better role here.

Conclusion:

A retrieval system is evaluated based on accuracy, efficiency, and overall utility. Accuracy is considered the most important aspect in the retrieval research world. We cannot easily determine the accuracy in the documents in the real world. It is not easy to implement manually on every document. It is dependent on the implementation of the algorithm. Each method has its own pros and cons. We must decide the approach which will work better depending on the scenario. From my review I want to state that in future, we should keep working on enhancing the existing approach to rank the relevance to retrieve the better result.

Sources

- Chapter 2 2 Information Retrieval - College of Computing. www.cc.gatech.edu/~isbell/tutorials/TextRetrieval.fm.pdf.

- “A Survey of Text Mining: Retrieval, Extraction and Indexing Techniques.” CiteSeerX, International Journal Of Computational Engineering Research, 2012, https://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.412.6357&rep=rep1&type=pdf.
- “A Study on Information Retrieval Methods in Text Mining.” *Https://Www.ijert.org/Research/a-Study-on-Information-Retrieval-Methods-in-Text-Mining-IJERTCONV2IS15028.Pdf*, https://www.ijert.org/research/a-study-on-information-retrieval-methods-in-text-mining-IJERTCONV2IS15028.pdf
- “Overview of Text Retrieval.” *Times.cs.uiuc.edu*, http://times.cs.uiuc.edu/course/410/note/tr.html