Inverter Index CSIS 400

Michael Scanlon & Shah Hassan Syed

The system that we construct uses two different types of index. One without stopwords and one with stopwords. The construct time of Non-stopword index is significantly less than the construct time of stop-word index.

The reason behind it is the addition of stop words in the stop words index takes significantly more time, considering the vast amount of stop words in the documents.

Based on this fact, the system also showed a difference in the performance of query search, whether it be single query or batch query. When we turned the stop word search on, we saw a significant increase in the number of documents that were retrieved, compared to when we searched without stop words. Thus increasing the recall when searched through the stop word index, and less precision

For a single query “Is the disease of Poliomyelitis (polio) under control in the world?”, the number of documents retrieved, when searched through the non-stop words index was 4817, with Precision of 0.0000 uptil 100 documents.

Average Precision(non-interpolated) while searching through the non stop word index

Query: “Is the disease of Poliomyelitis (polio) under control in the world?”

|  |  |
| --- | --- |
| Num. of Docs | Precision |
| 5 | 0.0000 |
| 10 | 0.0000 |
| 15 | 0.0000 |
| 20 | 0.0000 |
| 30 | 0.0000 |
| 100 | 0.0000 |
| 200 | 0.0050 |
| 500 | 0.0040 |
| 1000 | 0.0020 |

On the other hand, searching the same query through the stop word index retrieved 19830 documents, compared to the 4817 of the non-stop word search.

Since there was a high increase in recall in this search, the precision fell at 200 docs from 0.0050 to 0.0000 and at 500 docs from 0.0040 to 0.0000.

The Precision averages for interpolated precision, also fell down from 0.0099 to 0.0023, when changed the search from non-stop word to stop word.

Average Precision(non-interpolated) while searching through the stop word index

Query: “Is the disease of Poliomyelitis (polio) under control in the world?”

|  |  |
| --- | --- |
| Num. of Docs | Precision |
| 5 | 0.0000 |
| 10 | 0.0000 |
| 15 | 0.0000 |
| 20 | 0.0000 |
| 30 | 0.0000 |
| 100 | 0.0000 |
| 200 | 0.0000 |
| 500 | 0.0000 |
| 1000 | 0.0020 |

Compared to the boolean search vector Search gave a massive increase in the precision, on all documents. Since vector search was modified to only return 1000 documents, all the queries returned 1000 documents.

Compared to the boolean search, the interpolated precision of the vector model was 1.000 at all intervals compared to the interpolated precision of 0.0099 for non-stopword boolean search and 0.0023 for stop word boolean search.

The same query that we have mentioned above gave the following results:

Average Precision(non-interpolated) while searching through vector model

Query: “Is the disease of Poliomyelitis (polio) under control in the world?”

|  |  |
| --- | --- |
| Num. of Docs | Precision |
| 5 | 0.4000 |
| 10 | 0.2000 |
| 15 | 0.1333 |
| 20 | 0.1000 |
| 30 | 0.0677 |
| 100 | 0.0200 |
| 200 | 0.0100 |
| 500 | 0.0040 |
| 1000 | 0.0020 |

We can see the significant increase in the precision when we used the vector model, compared to the boolean model.

The query “Determine the usefulness and effectiveness of continuing to maintain export controls on encryption software”, was one of the best performing query with an average precision of 0.77. The reason could be the connectedness between terms such as software and encryption, or export and control. And these are the core words.

The query, “Evidence that radio waves from radio towers or car phones affect

brain cancer occurrence?” was one of the worst performing queries in the vector model

With average precision of 0.024 on all relevant documents. A reason could be that terms such as radio, car and phone are used in many different contexts, rather than just brain cancer. A possible improvement could be using bigrams. Words such as Radio and towers can exist individually in many different documents. However the bi-gram “radio tower” is very unique and narrows down the retrieved documents,