**Reading Summary**

**Shuning Chen**

**Andrew ID: shuningc**

**Chapter 1,5.1**

In chapter 1, we learnt about some basic ideas for information retrieval.

Ad-hoc retrieval is the original standard retrieval way in that users who require the information provide the queries directly.

Boolean retrieval model is built for searching. Usually, operators such as “AND” and “OR” are applied between terms. Binary digits show whether a term appears in a document or not. Inverted lists come up as to reflect the relationship between terms and documents. One term is linked to a posting list. Posting lists are composed of the document IDs. They are unique markers for identifying a document. The term key may not only contain the word but also information like frequency. These terms are the words split in documents. They are valid words after tokenization. Once a term is seen shown in a document, the document ID will be written down and linked to the list. The list can be realized by singly linked lists or dynamic sized arrays.

The more effectively the users can get documents required, the higher efficiency the search achieves. The efficiency is measured by the relevance and recall of the documents. Relevance means how many documents are relevant to the topic in the searching result. Recall means how many documents would be returned in the relevant ones.

In institutes like Westlaw, boolean retrieval model is preferred for it is more precise. We need to keep a balance between relevance and recall according to our search requirements. We can get more precise results. The relevance is high but recall is low. If we use “OR” instead of “AND”, the recall is higher and the relevance is lower. Ranked retrieval is a sort of ambiguous search. The result is produced by comparing which documents match the queries best rather than determine which documents is required directly. Different terms get different scores based on the times they appear in documents. The results are ordered according to terms’ importance.

In chapter 5.1, two laws are introduced. One is Heaps’ law. It states that the number of terms has an exponential relationship with the the tokens in documents. K is a constant, ranging from 30 to 100. b is power, ranging from 0.4 to 0.6. The prediction result is very close to the real occurrences. Another law is Zipf’s law. It states that a term’s frequency is inversely proportional to its ranking. The most frequently appearing term takes about half account of the total words. The frequency distribution is relatively more accurate compared to the prediction at two tails.