05.02.2013

CS 533 – Information Retrieval Systems

Prepared by: Mehmet Ali Abbasoğlu

Week No. 1

- Users query Information Retrieval Systems (IRS) to get their needed relevant documents.
- Information Retrieval Systems are required to provide most relevant documents in a ranked way.
- To achieve the most optimal results IRS apply
  - **Clustering**, to group query results together, and detect similar results.
  - **Diversification,** to find results in different domains to ad-hoc queries.

    Such as returning result documents in domains like “animals”, “cars”, “operating systems” and “cocktail” for query about “jaguar”
  - **Ranking,** to provide most important and most relevant results at top.

Implementation of Information Retrieval System

- Information retrieval systems can be implemented in 3 basic steps:
  1. Documents are loaded into information retrieval system.
  2. Stop-words are cleaned from the documents.
  3. Documents are indexed with the help of several methodologies.

- As a part of indexing “Document Term Matrix” can be build from the documents. “Document Term Matrix” includes term frequency information in a matrix representation. In matrix, terms are placed in columns and documents are placed in rows. The value for the specific place in the matrix shows that how many times corresponding term appears in the corresponding document.

An example Document Term Matrix is

$$D = \begin{array}{cc} \left[\begin{array}{ccccccc} \; & \; & \; & \; & \; & \; & \; \\ & & & & & & \\ & & & & & & \\ & & & & & & \\ & & & & & & \\ & & & & & & \\ & & & & & & \end{array}\right] & \begin{array}{c} d_1 \\ d_2 \\ d_3 \\ . \\ . \\ . \\ d_n \end{array} \\ \begin{array}{ccccccc} t_1 & t_2 & t_3 & . & . & . & t_n \end{array} & \end{array}$$

In document term matrices, instead of specifying number of occurrences, they can reverted into their binary form. Binary form document term matrices build from 0s and 1s, which are indicating that term is appearing once or not in the corresponding document.

Stop-words are words which are filtered out prior to processing of natural language text. Any group of words can be chosen as the stop words for a given purpose. Most common stop-words can be listed as "a, an, by, in, of, overview, systems, the"

For example purposes lets use following texts as documents.

Doc1 : "Information Retrieval by Parallel Document Ranking"

Doc2 : "An Analysis of Parallel Text Retrieval Systems"

Doc3 : "Information Retrieval in the Law Office: An Overview"

The underlined words are eliminated before building document term matrix, because they are listed in the stop-word list before.

After eliminating stop-words, we can list indexing terms as follows:

1. Analysis
2. Documents
3. Information
4. Law
5. Office
6. Parallel
7. Ranking
8. Retrieval
9. Text

We can build our Document Term Matrix with the listed indexing terms as

$$D = \begin{array}{cccccccccc} t_1 & t_2 & t_3 & t_4 & t_5 & t_6 & t_7 & t_8 & t_9 & \\ 0 & 1 & 1 & 0 & 0 & 1 & 1 & 1 & 0 & d_1 \\ 1 & 0 & 0 & 0 & 0 & 1 & 0 & 1 & 1 & d_2 \\ 0 & 0 & 1 & 1 & 1 & 0 & 0 & 1 & 0 & d_3 \end{array}$$

Tokens, Types and Terms

Text : “to sleep perchance to dream”

- A token is an instance of a sequence of characters in some particular document that are grouped together as a useful semantic unit for processing.
  - Ex: “to”, “sleep”, “perchance”, “to”, “dream”
- A type is the class of all tokens containing the same character sequence
  - Ex: “to”, “sleep”, “perchance”, “dream”
- A term is a (perhaps normalized) type that is included in the IR system's dictionary.
  - Ex: “sleep”, “perchance”, “dream”

There is a formula of number of types for n tokens like:

$$V_R(n) = K\ n^B$$

In that formula $V_R(n)$ returns the number of types while B is number of title.

Questions & Answers

1. What is diversification in the domain of information retrieval? Explain it with examples?

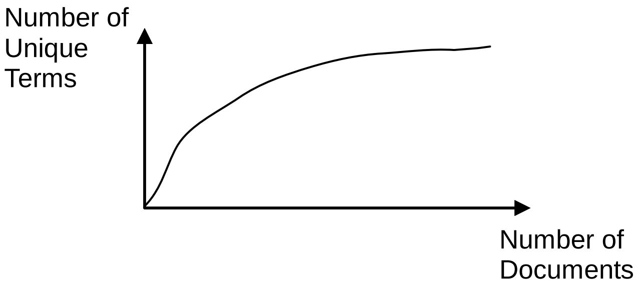
   Answer : Diversification is the practice of varying results while responding the query. For example for a query that asks for information about "Barcelona" should not only contain just for football club or city in Spain. There should be results from both domain.

2. What is the difference between document term matrix and binary representation of document term matrix?

   Answer: Document term matrix is a mathematical matrix that describes the frequency of terms that occur in a collection of documents. The origin of that matrix includes the occurrence numbers of terms. The binary representation includes just 0s and 1s, that are indicate term appear once in a document or not.

3. Explain the relation between "number of unique terms" and "number of documents". Use graph to explain the relation.

   Answer: Increase in the number of documents would increase number of unique terms until some point, but the amount of increase would be slowed from that point and after.



4. Build binary representation of document matrix for the following documents.

D1 : "new home sales top forecasts"

D2 : "home sales rise in July"

D3 : "increase in home sales in July"

D4 : "July new home sales rise"

Answer:

- Stop-words : “in”
- Terms :
  - t1 : “new”
  - t2 : “home”
  - t3 : “sale”
  - t4 : “top”
  - t5 : “forecast”
  - t6 : “rise”
  - t7 : “July”
  - t8 : “increase”

$$\begin{array}{cccccccccc} & t_1 & t_2 & t_3 & t_4 & t_5 & t_6 & t_7 & t_8 & \\ \left[\right. & 1 & 1 & 1 & 1 & 1 & 0 & 0 & 0 & \left.\right] d_1 \\ & 1 & 0 & 1 & 0 & 0 & 1 & 1 & 0 & d_2 \\ & 0 & 1 & 1 & 0 & 0 & 0 & 1 & 1 & d_3 \\ & 1 & 1 & 1 & 0 & 0 & 1 & 0 & 1 & d_4 \end{array}$$

5. Please write how many tokens and how many types existing in the following text.

“Information retrieval is the activity of obtaining information resources relevant to an information need from a collection of information resources. Searches can be based on meta-data or on full-text indexing.”

Answer : number of token : 30

number of types : 23