Passage Based Retrieval

(COSC 488)
Nazli Goharian
nazli@cs.georgetown.edu

Motivation:
• Only small section of a relevant document contains the information relevant to the query. Example: book chapter.
• Non-relevant sections may mask the relevant segment causing a lower relevance ranking for that document.
Passage Based Retrieval
(Algorithm)

• Identify document sections (passages) – various approaches exist

• Measure the similarity of each passage to a query

• Merge the passages’ similarity measures – various approaches exist

Example:
– Document D₁
– Sections of D₁: S₁, S₂, S₃, S₄, Sₙ
➢ Instead of calculating SC(D₁,Q), calculate:
  SC(Sᵢ,Q), for i=1,n

Then, merge similarity measures SC(Sᵢ,Q)
Identify Passages: Marker-based Passages

- Using section headers or paragraphs
- The passages are bounded to certain number of terms to avoid too long or too short sections.
  - Partitioning long passages; gluing short passages
  - Sample algorithms: discourse, window ([non]overlapping)
- Little improvement in accuracy
- Problem:
  - Multiple concepts in one section (caused by: author’s choice; combing short passages)
  - Not a good semantic partitioning

Discourse Passage (DP)

- Discourse passages are based on logical components such as discourse boundaries like a sentence

The sky is blue. How beautiful! It was cloudy yesterday.
Non-Overlapping Window Passage (NWP)

- Window based passage approach defines a passage as $n$ number of words

The sky is blue. However, it is raining continuously since morning.

Overlapping Window Passage (OWP)

- Document is divided into passages of evenly sized blocks by overlapping $n/2$ from the prior passage and $n/2$ from the next passage.

The sky is blue. However, it is raining continuously since morning.
Identify Passages:
Dynamic Passage Partitioning

- Find automatically good partitions based on the particular query.

- Sample algorithm:
  - Find query term $t_i$ in document $D_i$
  - Build passage from location of $t_i$, $n$ to $n+p$ ($p$ is a variable passage size)
  - The next passage starts from $n+(p/2)$ to overlap with previous passage to avoid splitting sections

Merging Passage-based Similarity Measures

- More than twenty different methods
- Ranking the SC of passages of $D_i$
- Combine document level SC with SC of highest rank passage
Summary
(Passage-based Retrieval)

• Popular for very large documents (such as book, congressional record,…) – makes the search results meaningful

• Useful to perform text mining & analysis on portions of data