

# Passage Based Retrieval

(COSC 488)

Nazli Goharian

nazli@cs.georgetown.edu

1

# Passage Based Retrieval

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.

2

## 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

3

## Passage Based Retrieval

- Example:
  - Document  $D_1$
  - Sections of  $D_1$ :  $S_1, S_2, S_3, S_4, S_n$
  - Instead of calculating  $SC(D_1, Q)$ , calculate:  
 $SC(S_i, Q)$ , for  $i=1, n$
  - Then, merge similarity measures  $SC(S_i, Q)$

4

## 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

5

## 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.



6

## 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.



7

## 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.



8

## Identify Passages: Dynamic Passage Partitioning

- Find automatically good partitions based on the particular query.
- Sample algorithm:
  - Find query term  $t_j$  in document  $D_i$
  - Build passage from location of  $t_j, n$  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

9

## 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

10

## 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