IR seminar, 2021
Result Snippets

By Kris Farrugia and Eva Elžbieta Sventickaitė
Agenda

• What are Result Snippets?
• Advantages of Query Biased Summaries in Information Retrieval
• Result Snippet in Web Search
• Q&A
Google defines a snippet as “a description of or an excerpt from the webpage” that follows the title and precedes the URL and Cached link. Simply put, the snippet refers to the description portion of a Google search listing. 2010-03-18

https://searchengineland.com › anatomy-of-a-google-snippet...

Anatomy Of A Google Snippet - Search Engine Land

Ieškoti: What are snippets in Google?
What are Result Snippets?

Snippet:
  • title + several sentences + {could be also URL, metadata}

Static vs Dynamic

How?:
  • rank documents against a query, return result list

Why?:
  • more efficient information retrieval method
Advantages of Query Biased Summaries in Information Retrieval

Authors and Date

• Anastasios Tombros and Mark Sanderson in 1998

What was the goal?

• To investigate the utility of document summarisation, specifically in query biased summaries.
How was the goal achieved?

- Based upon the users’ **speed and accuracy** in identifying relevant documents
- Query biased vs. static predefined summary
Predefined Static Summary Example

727 results returned, ranked by relevance.

Landings: Aircraft Manufacturers  91%  [find similar]
http://www.landings.com/_landings/pages/aircraft-manuf.html

Landings: Avionics  90%  [find similar]
Digifly USA Icc aviation electronic instruments GPS moving map, flight and engine instruments. Skylight Avionics Manufacturer of electronics interfaces and cockpit information displays for business and commuter aircraft.
http://www.landings.com/_landings/pages/avionics.html
Automatic Generation Summarisation System

• An automatically generated summary can minimise users’ need to refer to the full text and provide enough information to support their decisions.

• The new approach presented provides a more complete approach which significantly improves both the accuracy and speed of user relevance judgements.
Summarization System

• It was based on a number of sentence extraction methods (Paice 1990) that utilise information both from the documents of the collection and from the queries used.
• Documents to be summarised: Articles from the WSJ taken from the TREC collection.
• Four sentence extraction methods were implemented.
Sentence Extraction Methods

1. Title Method
2. Location Method
3. Term Occurrence Information
4. Biasing summaries towards queries

• Summing scores up yielding the summary for each document (summary, 15% of doc’s length, max. 5 sentences)
1. Title Method

- Title --> the major subject of the article and preview
- Terms occurring in the title section of the documents were assigned a positive weight
2. Location Method

• The leading paragraph of each article provides much information on the article’s content
• An ordinal weight was assigned to the first two sentences of each article
• A positive ‘heading score’ was assigned to each of the sentences comprising a heading
3. Term Occurrence Information

- The number of TOs to assign weight of sentences instead of assigning weight to each term according to its frequency
- TO value: 7 for medium-sized documents (25-40 sentences), less or more depending on size
- Cluster of terms added to notion of significance based upon their location and lexical semantic (relatedness in meaning and their distance apart)
4. Biasing summaries towards queries

- ‘Query score’ was calculated for each sentence based on the distribution of query terms in each sentence.
- The larger the number of query terms in a sentence, the more likely that sentence conveys a significant amount of the information need expressed in the query.
Experimental Design

Design Considerations:

- Experimental conditions (two levels: (i) static predefined summary or (ii) query-biased summary)
- Groups of subjects (10 subjects to each level)
- Situational Variables (background noise, subjects’ behaviour, etc)
Experimental Design

• Retrieval Task (To identify as many relevant documents as possible)
• Queries used (Randomly select from TREC collection)
• IR system used (To generate the retrieved document lists: classic document system + tf-idf + stop words removal + word stemming)
Findings - Recall and precision

Recall

- Subjects using the summaries: 65.6%
- Subjects using a typical IR output: 49.76%
+15.84%

Precision

- Subjects using the summaries: 55.32%
- Subjects using a typical IR output: 44.29%
+13.03%
Findings - Speed

Small difference however amounting to a 13% increase in the average number of documents examined.

![Diagram showing average number of documents examined with summaries and a typical IR output with 22.62 vs 20 documents respectively.]
Findings - Reference to the full text

- Subjects using the summaries: 1.32%
- Subjects using a typical IR output: 23.7%
Conclusions

• Query biased summaries assist users in finding relevant documents more accurately and more quickly.
• Reduce the number of times the user needs to refer to the whole text
• Brief and informative (related to query) summaries
Who and when?

- A. Turpin, Y. Tsegay, D. Hawking, H. E. Williams in 2007

What was the goal?

- find algorithms and data structures for more efficient query biased snipped generation
Result Snippet in Web Search

Why are snippets important?
  • Direct answer + 10 snippets per search → computational load

The solution:
  ...new algorithm and compact single-file structure.
Methodology

What was done and how?

- Simulation
- Reported:
  - Proportion of time spent for disk seeks, disk reads and CPU processing.
- Baseline:
  - Simply compressed documents (zlib),
- Data:
  - TREC Web Track (wt10G, wt50G and wt100G)

...limitations?
Rank calculation

TEXT, w: word, q: query

- l: line
- h: heading
- c: w, which are q, repetitions
- d: distinct q terms that match w
- s: score (weight)

\[ s = [l*h*c*d*s*k] \]

max weight sentences → snippet
CTS - Compressed Token System

Semi-static compression method:

- Words and non-words to integers.
- Using ESCAPE for rare words.
- Two move-to-front queues are kept.
- When filled and new found: old discarded.

Sentences can be scored without decompressing document → decode only highest ranking.
Sentence reordering
## Findings

### Time

<table>
<thead>
<tr>
<th></th>
<th>WT10G</th>
<th>WT50G</th>
<th>WT100G</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline</td>
<td>75</td>
<td>157</td>
<td>183</td>
</tr>
<tr>
<td>CTS</td>
<td>38</td>
<td>70</td>
<td>77</td>
</tr>
<tr>
<td>Reduction in time</td>
<td>49%</td>
<td>56%</td>
<td>58%</td>
</tr>
</tbody>
</table>

### Storage

<table>
<thead>
<tr>
<th></th>
<th>WT10G</th>
<th>WT50G</th>
<th>WT100G</th>
</tr>
</thead>
<tbody>
<tr>
<td>No. Docs. ($\times 10^6$)</td>
<td>1.7</td>
<td>10.1</td>
<td>18.5</td>
</tr>
<tr>
<td>Raw Text</td>
<td>10,522</td>
<td>56,684</td>
<td>102,833</td>
</tr>
<tr>
<td>Baseline(zlib)</td>
<td>2,568 (24%)</td>
<td>10,940 (19%)</td>
<td>19,252 (19%)</td>
</tr>
<tr>
<td>CTS</td>
<td>2,722 (26%)</td>
<td>12,010 (21%)</td>
<td>22,269 (22%)</td>
</tr>
</tbody>
</table>

If as little as 1% of the documents can be cached in RAM as part of the Snippet Engine, then around 75% of seeks can be avoided
Takeaways

• Result snippets help users as well as search engines.
• Web queries repeat, small number of queries make up a large volume of searches → caching.
• Caching reduces seek time.
• More compressed → more fit in cache → more seeks avoided → less load.
If Q then A

Else → keywords to remember:

1. Result snippet, Static vs Dynamic
2. Query Biased Summaries (Query Dependent)
3. Snippet Engine, Compressed Token System (CTS)