UNIVERSITI MALAYA

WIE3004

INFORMATION RETRIEVAL

SEMESTER 2 SESSION 2022/2023

CASE STUDY

PREPARED BY

WAN SURAYA BINTI WAN MOHD LOTFI

U2005345

PREPARED FOR

PROF. DR. SRI DEVI RAVANA

Question 1

Google Search is an example of a question-answering system based on information retrieval. The Google search engine uses cutting-edge information retrieval methods to quickly find answers from a large database of websites and other online sources. When a user types a question or query, Google's system analyses what is entered and searches through its database of web sites to locate pertinent short text passages or segments that offer precise answers. Say a user asks, "What is the capital of Japan?" as an example. The query will be processed by Google's question-answering system, which recognises that the user is looking for information about the capital of Japan. The system will retrieve information from trusted sources and identifies a snippet or passage that directly answers the query. In this case, it might extract the answer “Tokyo”. Google’s question answering system prominently displays this answer at the top of the search results page, along with the information source, allowing users to easily access the answers to their query without having to navigate through individual websites.

Question 2

2a) Google's question answering system employs several information retrieval models to provide accurate and relevant answers to user queries. The purpose these models is to provide direct and specific answers to user queries, rather than returning a list of relevant documents or web pages. Instead of expecting users to search through search results and extract the information they need, a Question Answering system aims to deliver concise and accurate answers in a more intuitive and efficient manner.

The process of an Information Retrieval based Question Answering system like google involves indexing, query processing, document retrieval, passage retrieval and finally answer processing. Information retrieval models are typically employed during the document retrieval process of a question answering system. This process involves searching through a collection of documents to identify those that are relevant to the given question.

Google employs various retrieval models, and the models include both traditional information retrieval techniques and more advanced machine learning-based methods. Some common retrieval models used by Google include the vector space model, and language models like BERT.

- Vector Space Model

  This model allows for vector similarity search, also known as "nearest neighbour search," by representing documents and queries as vectors in a high-dimensional space. The scope of traditional keyword-based searches and their handling of synonyms are both constrained. In contrast, vector search uses vectors to represent and search content, with each vector defining similarity to specific topics. Google can quickly locate relevant material for users worldwide through comparing the distances and similarities across vectors while taking into account many meanings, synonyms, and entities linked together.

  Binary options are used in keyword search, which makes it difficult to convey the true meaning of the content being searched. Vector search allows for more precise content findings since vectors capture subtle nuances and meanings. A subset of content having certain properties, such as "much about actors, some about movies, and a little about music," can be represented as a vector. Additionally, they offer flexibility in representing previously unidentified or undefinable groupings. This is especially useful for identifying emergent content categories that could be difficult for humans to foresee or manually update in large databases.

  All in all, Google's system uses vector similarity search to provide a more sophisticated and thorough approach to analysing and retrieving relevant content, taking into account numerous interpretations and capturing complex meanings in a scalable and efficient manner.

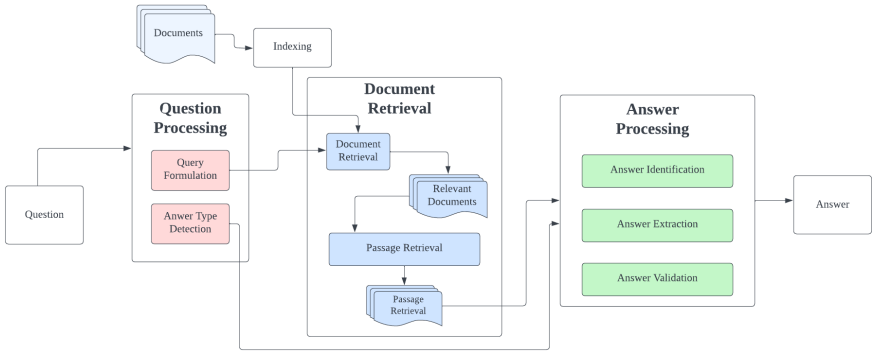
- Language Model

  To deliver accurate results, the Google Search engine tries to comprehend language. It is difficult for computers to comprehend context, tone, and intention. The Bidirectional Encoder Representations from Transformers (BERT) paradigm, which aids in processing language and comprehending its context, is used in Google's method. BERT considers the smaller words and their order, enabling a more nuanced comprehension of queries. The model is trained by anticipating phrases that are

missing words, which enables it to understand the meaning of various words in diverse circumstances.

Overall, Google's question answering system combines information retrieval models, query understanding, and ranking algorithms to provide accurate and relevant answers to user queries. The models aim to match queries with relevant documents, understand the context and intent of the queries, and extract answers from the retrieved information.

Question 2b)



The design of a question answering system has a specific vital component. There are three distinct modules used in a question-answering system, Question Processing Module, Document Retrieval Module and Answer Processing module.

1. Question Processing:

   When users enter their queries into Google's search engine, it initiates the question processing phase which involves 2 tasks:

   - Query Formulation
     1. Google employs a wide range of techniques to pre-process the query, including tokenization, stop word removal, and language analysis. These processes help normalize and clean the question text, making it easier to understand and process further.
     2. For instance, “What are the best hiking trails in Malaysia?” is then formulated into a structured query like “best hiking trails Malaysia”
     3. By structuring the question in this way, the search engine is more likely to recognise the essential components of the user's information requirements and deliver more precise and pertinent search results. Finding the needed information is a greater possibility when the user's question is refined, and the search is accurate thanks to query formulation.

- Answer Type Detection
    1. Natural Language Processing (NLP) methods are utilized to analyze the question's syntax, structure, and semantics. This allows Google's systems to understand the intent behind the query, extract relevant information, and identify key entities or concepts.
    2. For instance, if the query contains terms like "how to," it implies that the user is seeking for a step-by-step instruction manual or guide. Google can customise the search results to give the user more pertinent and helpful information by identifying the answer type.

2. Document Retrieval:

   The document retrieval module accepts the formulated question as its input. Google's search engine has indexed an enormous collection of web pages from across the internet, comprising billions of documents. Information retrieval models, including Google's proprietary algorithms such as RankBrain, Neural Matching and BERT come into play during the document retrieval stage. The module is divided into major tasks:

    - The retrieval process involves matching the user's query against the indexed documents and identifying the most relevant web pages for further analysis.
    - The documents will be sorted according to their similarity and relevance to the question.
    - Filter the set of documents obtained from previous step and divide the documents into smaller units such as sentences or paragraphs that are most likely to produce or contain a concise answer, then turn them into passages.
    - Order the passages by similarity and relevance to the question.

3. Answer Processing:

   Once the list of passages by the previous module are retrieved, Google's question answering system proceeds with the last module, answer processing. The answer processing module considers the set of passages and performs three major tasks:

    - Identify statements/answers within the set of passages.
    - Extract the relevant answer by choosing the appropriate phrases and words that answer the question. Techniques such as natural language understanding are employed to analyse the content of the retrieved web pages and extract potential answers.
    - To grasp the textual material contained in the web pages, advanced algorithms and machine learning models are used, recognising passages or

sections that provide relevant answers. Answer ranking and validation mechanisms are applied to determine the most accurate and relevant responses. Factors considered may include the credibility and quality of the information source, relevance signals from user interactions, and the context of the user's query.

4. Answer Presentation:

   Users are given the chosen answers by Google in an approachable manner on the search engine results page. The answer may show up in the search results as a featured snippet, a direct answer box, or a prominent result depending on the nature of the query and the information that is available. The goal of Google's systems is to deliver concise, accurate responses that immediately respond to user inquiries. To help the user comprehend and encourage further inquiry, additional context, related details, and sources might be included alongside the answer.

Question 2c)

Yes, the chosen Question Answering system, Google Search, combines artificial intelligence (AI) with information retrieval (IR) techniques to enhance its functionality. The first AI google used into search is called RankBrain. With the aid of RankBrain, Google was able to better identify the relationships between broad queries and concepts from the real world. It basically helps google rank search results and is part of the ranking algorithm.

To give an example of how Google's RankBrain is used, if you search for “How to repaire a broken heart”, Google’s systems learn from similar past queries and associated results and understands that this query involves emotional healing rather than a physical repair. Therefore, it provides search results on coping mechanisms, therapy and similar solutions for moving forward after a breakup or emotional trauma.

In addition to RankBrain, Google introduced Neural Matching to further enhance its AI for search. Neural Matching helps Google understand the connection between queries and pages by analyzing the entire query or content on the page. It interprets the information within the context of the page or query. For example, "why does my TV look strange" will be understood by Google, who connects it to the idea of "The soap opera effect." Even if the exact terms aren't used in this situation, Google is still able to retrieve pages discussing the soap opera effect.

Another utilized AI technique by google is called BERT (Bidirectional Encoder Representations from Transformers) for pre-training natural language processing. BERT helps Google understand the many meanings and intentions expressed by word combinations, including the order of words on a page. As a result, even seemingly unimportant terms in requests are considered.

Here's an illustration: Before BERT, Google's understanding of queries like "symptoms of a chest cold in infants" was limited to individual keywords, lacking full context and intent comprehension. With BERT, Google's algorithm recognizes the specific focus on chest cold symptoms in babies, providing more precise and tailored search results for infant healthcare needs.

Question 2d)

In my opinion, the most prominent challenges faced by information retrieval-based Question Answering Systems would be contextual reasoning. Many questions rely heavily on contextual understanding, where the answer depends on the surrounding context. However, understanding context from a passage or document can be challenging. Therefore, techniques such as pre-training models on large-scale datasets, such as BERT (Bidirectional Encoder Representations from Transformers), have been proposed to enhance the system's ability to capture and reason with contextual information. These models can learn contextual representations that capture the nuances and dependencies within a passage, facilitating more accurate answers.

Another challenge is the lexical gap which is caused when user may use different words, synonyms, or phrases to describe the same concept or entity. For example, a user might search “automobile” while the relevant information source or the question answering system may use the term “car”. This difference in terminology can lead to mismatches and hinder the retrieval of accurate and relevant results. In order to address this challenge, it is required to use techniques that can effectively handle synonymy, polysemy and context. Natural Language Processing methods such as query expansion and reformulation can be used to bridge the gap by incorporating a broader range of relevant terms and synonyms.

Refereces

Google. (n.d.). *In-depth guide to how Google Search works*. Google. https://developers.google.com/search/docs/fundamentals/how-search-works

Hamza, A., En-Nahnahi, N., Zidani, K. A., & El Alaoui Ouatik, S. (2021). An Arabic question classification method based on new taxonomy and continuous distributed representation of words. *Journal of King Saud University - Computer and Information Sciences*, *33*(2), 218–224. https://doi.org/10.1016/j.jksuci.2019.01.001

Nakayama, H. (2020, July 9). *Building an application of question answering system from scratch*. Medium. https://towardsdatascience.com/building-an-application-of-question-answering-system-from-scratch-2dfc53f760aa

Roy, J. (n.d.). *Everything you need to know about Google BERT*. SEO Site Checkup. https://seositecheckup.com/articles/everything-you-need-to-know-about-google-bert

Sato, K., & Chikanaga, T. (n.d.). *Find anything blazingly fast with Google's vector search technology | google cloud blog*. Google. https://cloud.google.com/blog/topics/developers-practitioners/find-anything-blazingly-fast-googles-vector-search-technology

Schwartz, B. (2022a, February 18). *Video on how Bert helps with Google Search - it's the little words*. seroundtable.com. https://www.seroundtable.com/video-how-bert-helps-with-google-search-32946.html#:~:text=BERT%20is%20a%20machine%2Dlearning,it%20thought%20were%20most%20important.

Schwartz, B. (2022b, October 24). *How google uses artificial intelligence in google search*. Search Engine Land. https://searchengineland.com/how-google-uses-artificial-intelligence-in-google-search-379746

Slawski, B., Robinson, J., Simmons, G., Gurung, S., Santos, J., Eric, Heaton, A., Ede, P., Keadle, L., Carolm, Gladstein, B., Ahmad, Sells, M., K, C., Fox, J., Cobe, M., Paul, Viau, P., Baptiste, … Carter, D. (2020, January 19). *How google might generate search results snippets*. SEO by the Sea ⚓. https://www.seobythesea.com/2013/02/google-snippets-search-results/

Taylor, D. (2022, October 13). *Information retrieval: An introduction for seos*. Search Engine Journal. https://www.searchenginejournal.com/information-retrieval-seo/464164/#:~:text=Vector%20Space%20Models,queries%20are%20represented%20as%20vectors.