UNIVERSITI
MALAYA

# WIE3004

## INFORMATION RETRIEVAL
## SEMESTER 2 SESSION 2022/2023

## FINAL ASSIGNMENT

## PREPARED BY

WAN SURAYA BINTI WAN MOHD LOTFI
U2005345

## PREPARED FOR

PROF. DR. SRI DEVI RAVANA

1. Bidang Capaian Maklumat telah berkembang dengan pesat pada dekad yang lalu. Senaraikan TIGA (3) perkembangan atau bidang penyelidikan yang berlainan di bawah bidang Capaian Maklumat. Terangkan kepentingan perkembangan ini.

   *The Information Retrieval field has evolved rapidly during the last decade. List* ***THREE (3)*** *different developments or research area under the field of Information Retrieval. Explain the importance of these developments.*

[6 markah/*marks*]

*1)* Multimedia in Information Retrieval

Everyone uses multimedia in their daily lives; it is the state-of-the-art in information and digital media representation. Traditional information retrieval methods are often text-based, making them ineffective for finding and retrieving multimedia content. By April 2022, 66,000 photos were shared on Instagram, 500 hours of video were uploaded to YouTube, 2,430,000 Snaps were shared on SnapChat, 1,700,000 pieces of multimedia content were posted on Facebook, and 231,400,000 emails containing media were sent, according to Stefan Wagenpfeil et al.'s 2023 study. The underlying infrastructure and information retrieval systems face challenges as a result of the ongoing growth of these huge volumes. Because of this, there is now an emphasis on creating techniques for content-based image and video retrieval, audio analysis, and multimodal fusion in order to increase the retrieval efficiency and accuracy for multimedia data. The importance of developing methods to efficiently search for and retrieve pertinent multimedia information is also reflected in the fact that multimedia content is becoming more and more important and accessible online.

*2)* Query Expansion Techniques

Hiteshwar et al., 2019 discuss how it is becoming difficult to find relevant material online using a query made up of a few keywords due to the web's constantly expanding size. An online search frequently produces unrelated results. This is due to a number of factors. First, the user-submitted keywords may relate to a variety of subjects. The search results are therefore not narrowly focused on the subject of interest. Second, the query might not adequately express the user's search requirements because it is too brief. Third, until the user sees the results, he or she is frequently unsure of what they are looking for. Even though the user is aware of what he is looking for, he is unsure of how to phrase the right inquiry. Here, Query Expansion Techniques are crucial in obtaining pertinent results in the previously mentioned scenarios.

By reformulating the original user question, expanding it with more relevant terms, and modifying the weights of the terms in the enlarged query, query expansion (QE) approaches help information retrieval systems improve the user's original query. In order to improve retrieval performance and acquire more relevant documents, these extra terms may also contain synonyms, similar terms, or terms taken from an ontology or thesaurus. The value of query expansion techniques rests in their capacity to solve the vocabulary mismatch issue, which arises when the user query may not contain words that exactly match those found in the relevant documents. These strategies improve the overall search experience for users by increasing the likelihood of finding more pertinent documents by expanding the query.

### *3)* Conversational Information Retrieval

According to Mohammad Aliannejadi et al., users frequently struggle to express their complex information requests in a single query. As a result, customers might have to search through numerous result pages or rephrase their searches, both of which can be a frustrating experience. Alternatively, systems can improve user satisfaction by proactively asking questions of the users to clarify their information needs. Asking clarifying questions is especially important in conversational systems since conversational systems can only produce a finite number of results.

Information retrieval through natural language conversations is the goal of conversational information retrieval. The strategies for comprehending user inquiries in conversational formats, producing pertinent responses, and keeping context during the dialogues are explored in this research area. Conversational information retrieval has grown in significance as a means of enabling more natural and interactive information-seeking experiences with the emergence of voice assistants, chatbots, and conversational interfaces. It enables users to conversely communicate their information needs, obtain prompt responses, and iteratively hone their searches, resulting in more effective and efficient information retrieval.

2. Bezakan enjin carian Yahoo dan Google dari segi fungsi, keberkesanan dan kecekapan. Jalankan penyelidikan anda sendiri.

   *Differentiate the search engines Yahoo and Google in terms of functionality, effectiveness, and efficiency. Conduct your own research.*

   [8 markah/*marks*]

In the realm of search engines, two prominent players have dominated the online landscape for decades. Both Google and Yahoo, two of the most popular and widely used search engines, are significant players in the computer software sector. These search engines have become synonymous with searching information on the internet, yet each one is distinctive in a number of ways. As each firm offers a significantly different user experience, both Google and Yahoo are different in their features, tools, and even search results.. To fully understand the distinctions between Yahoo and Google, it is important to explore their functionalities, effectiveness, and efficiency.

1) *Functionality*

A search engine's functionality covers all of its features and services. Both Yahoo and Google provide basic search functionality that enables users to input queries and get relevant outcomes. However, they also differ in what extra services and products they offer.

A fully automated search engine, Google Search uses software known as web crawlers to continuously search the web for pages to add to their database. In reality, the vast majority of the pages featured in our results aren't manually submitted for inclusion, but rather are discovered and added as our web crawlers browse the web.

Google Search works in three stages, crawling, indexing and serving search results. Crawling uses automated tools known as crawlers, Google retrieves text, photos, and videos from web pages it finds online. Indexing is where google scans a page’s text, photos, and video files before storing the material in its massive database, the Google index. Finally, serving search results is when a person conducts a search on google, google returns results that are relevant to their query.

On the other hand, Yahoo is a web service provider that offers both a search engine and a directory of World Wide Web pages organized in a hierarchy of topic categories. While the Yahoo web portal started off as a web directory, it soon added other services such as email, news, and finance.

In terms of ranking, "PageRank" is the priority ranking algorithm used by Google. Each web page is given a relevancy score by this algorithm, which is largely based on the frequency and placement of the keywords on the page. In order to rank the page, useful material must be presented on the webpage, and appropriate keywords must be used.

As for Yahoo, they utilises the web directory approach as a component of its ranking algorithm. Your website rises to the top as more people click on it. Since the page title is the primary determinant of ranking, the keywords must be present. In addition, it also ranks pages in order to provide relevant results using the page ranking methods that has been uploaded on the internet.

In terms of services, Google has a larger market share and a wider scope compared to Yahoo. It offers various services beyond search, including cloud computing, advertising technologies, software technologies, and a range of applications such as email, office suite, web browsing, social networking, and photo editing. Google also powers platforms like YouTube, Blogger, Google+, Google Chrome, and Google Maps. Additionally, Google developed the Android operating system and Google Chrome OS. It has also ventured into hardware with smartphones like Nexus. Google's primary revenue source is advertising, particularly through Google AdWords.

On the other hand, Yahoo has a smaller market share but provides services that Google cannot match. These include Yahoo Answers Q&A, Yahoo Finance, Flickr, backlink reporting, privacy features, and entertainment options. Yahoo's offerings encompass Yahoo Search, Yahoo Mail, Yahoo advertising, Yahoo News, Yahoo Groups, online mapping, fantasy sports, and social media. In summary, Google has a broader range of services and a larger market share, while Yahoo has its own unique offerings and strengths in areas such as entertainment and privacy.

2) *Effectiveness*

In terms of effectiveness of a search engine, this can be measured by using various metrics and evauation techniques. One of them being precision and recall. A study by Sumeer Gul et al, 2019 has evaluated the retrieval performance of Google and Yahoo on navigatioinal queries with the help of two widely used evaluative measures that is precision and relative recall in the field of life science and biomedicine.

Precision is a measure of the relevance of the retrieved results. It calculates the ratio of pertinent documents among the retrieved documents. Higher precision indicates a lower likelihood of retrieving irrelevant results. For example, if a search engine returns 10 results for a query, and 8 of them are relevant, the precision is 8/10 or 80%.

Recall, on the other hand, measures the completeness of the retrieved results. It calculates the ratio of pertinent documents retrieved compared to the total number of relevant documents in the collection. For example, if there are 100 relevant documents in the collection, and the search engine retrieves 80 of them, the recall is 80/100 or 80%.

At the end of the research by Sumeer Gul et al, 2019, he declares that Google accomplishes the highest precision for navigational queries and also the highest mean relative recall for navigational queries in comparison to Yahoo.

To test this study, I have searched “Kolej Kediaman Kelima Univerisiti Malaya” on both Yahoo and Google. Google’s search results directly shows the map and information for “Kolej Kediaman Kelima Universiti Malaya” and can be seen as a higher precision because the retrieved results directly match the specific query. On the other hand, Yahoo’s results providing facebook and other links, but not directly showing the desired map and information, may indicate lower recall since it did not capture all the relevant information that I was searching for.

In summary, Google has high precision because it accurately provided the specific information I wanted, and it also has high recall because it was able to retrieve and show the right information. On the other hand, Yahoo has lower precision because it did not provide the desired information, and it has lower recall because it missed retrieving the specific information I was looking for. Thus, I can conclude that there is a difference between the effectiveness of Yahoo and Google in terms of Navigational Queries.

Another research I did is searching “Prime Minister of Malaysia” on both search engines. Google provided me the accurate and recent one, which is Anwar Ibrahim while Yahoo still show random articles, links and images of the old prime minister instead of the current one. The variation can be due to differences in Yahoo and Google's search algorithms and underlying technologies. In order to provide accurate and current search results, Google has created sophisticated algorithms that take into account a variety of factors, including relevancy, information freshness, user intent, and semantic understanding. When possible, their search engine algorithm, which includes elements like knowledge graph, seeks to give users direct answers to their queries. In contrast, because Yahoo still favours old and well-established websites, their search algorithm is not as effective or up-to-date in finding the most relevant and recent information.

### 3) *Efficiency*

When discussing efficiency in search engines, there are numerous elements to take into account. One of them is the simplicity of use, which is a crucial efficiency aspect in deciding which search engine is the finest. According to Sumeer Gul et al., 2019, people prefer to use Google to conduct information searches because it offers a better user experience, more functionality, and ease of use. Knowledge Graph are one of Google features that make it easier for users to access and use the search engine. These capabilities allow users to access relevant information without the need for further clicks or navigation to external websites, which increases efficiency in finding the relevant information. On the other side, while Yahoo also offers a knowledge graph, the user interface is not as nice as Google. If you search for “Population of Malaysia” with both search engines, Google will provide you a display of an interactive graph which also links to a significant website while Yahoo returns a list of articles which is not a good user interface and increases the click on multiple search results to find the desired information. It also requires scanning and evaluation to determine the most relevant information, making it harder to efficiently find the desired information.

Another efficiency factor is the result retrieval efficiency. According to Irfan ul Haq Akhoon et al, 2019, he studied that Bing has higher result retrieval efficiency than Yahoo by exploring the total number of results retrieved from each search engine data collected with one search query. We can do similar research on Google and Yahoo with 5 one search query.

Table 1:
Number of Results Retrieved by Each Search Engine

| Query No | Search Term | Google | Yahoo |
|---|---|---|---|
| 1 | Books | 15,950,000,000 | 3,510,065,408 |
| 2 | Catalog | 2,470,000,000 | 2,500,000,000 |
| 3 | Circulation | 1,060,000,000 | 334,000,000 |
| 4 | Acquisition | 1,220,000,000 | 677,000,000 |
| 5 | Journals | 5,720,000,000 | 1,140,000,000 |
| Total | | 26,420,000,000 | 8,161,065,408 |

From Table 1, it is evident that the maximum numbers of results were retrieved by Google (26,420,000,000) followed by Yahoo (8,161,065,408) respectively. The study mentions how the explicatory results clearly shows that both search engines use different technology to find a particular web information. Thus, the table shows that Google has definitely higher result retrieval efficiency than Yahoo. The

The speed of retrieval is another aspect to take into account. Yahoo does not show the search speed, whereas Google does. However, Google typically returns results more quickly than other search engines, including Yahoo. In 0.19 seconds, it may present millions of results. This is due to their technical infrastructure, which is far stronger than that of other engines like Yahoo.

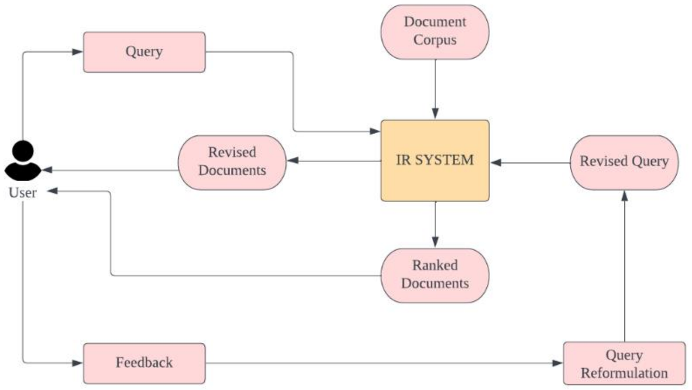
Although Google is renowned for its dominance and effectiveness as a search engine, depending on specific use cases or user preferences, there may be some components in Yahoo that are thought to be more efficient than Google. Yahoo Finance is one example. Yahoo offers a comprehensive and user-friendly tool for tracking investments and financial information. Real-time stock quotes, financial news, interactive charts, and other tools that investors find useful for keeping track of their portfolios are all provided.

Another is Yahoo News, which is the first result you see when entering the domain. By compiling news stories from multiple sources, Yahoo News gives users a centralised platform for staying up to date on the latest news. Users may quickly obtain pertinent news stories in their areas of interest thanks to its curated news categories and personalised content recommendations. On the other hand, Google does not offer an instant overview of news. Instead, consumers must conduct a specific search in order to obtain the news material they desire. This implies that readers may need to enter their search queries into Google in order to locate recent and updated news stories.

3. Lukis gambarajah aliran untuk model pemprosesan pertanyaan maklum balas relevan. Pada pendapat anda, bagaimana maklum balas relevan akan meningkatkan keberkesanan capaian maklumat?

*Draw the flow diagram for relevance feedback query processing model. In your opinion, how relevance feedback will improve the information retrieval effectiveness?*

[4 markah/*marks*]



In my opinion, relevance feedback can significantly boost the effectiveness of information retrieval systems as it is a technique that allows users to provide feedback on the relevance of the retrieved documents, which is then utilised to refine and improve subsequent searches.

With relevance feedback, iterative improvement is performed to increase the effectiveness of information retrieval. Information retrieval can be improved iteratively thanks to relevance feedback. Users can review the initially retrieved documents, give feedback, and use the revised query to start further searches. The system takes user feedback into account with each iteration and modifies the search approach accordingly. The accuracy and relevancy of the search results are continually enhanced over time by this iterative process of query improvement and feedback from users.

Additionally, they can assist the system in understanding user preferences. The IR system would be better able to comprehend user preferences and relevance criteria with the aid of relevance feedback. The system can learn about users' preferences, needs, and interests by examining the feedback they provide. By using this information, search results can be personalised and tailored to reflect the user's preferences more closely. The documents that are retrieved as a result are more likely to match the user's particular needs, increasing the effectiveness of discovering relevant information.

Therefore, relevance feedback definitely plays a crucial role in improving the effectiveness of information retrieval as it has several approaches like iterative improvement and learning user preferences, which plays an important role in achieving effectiveness.

4. Takrifkan terma-terma di bawah dan terangkan kepentingannya dalam capaian maklumat.

i. Pengurai ii.
Analisis pautan

*Define the terms below and explain its importance in information retrieval.*

*i. Parser*

*ii. Link Analysis*

[4 markah/*marks*]

i) A parser is a component or program that processes a sequence of text tokens in a document to recognize and interpret its structural elements. These structural elements can include titles, links, headings, and other components that define the organization and layout of the document.

To begin with, a tokenizer is used within the parser to identify individual "words" or tokens in the text. The tokenizer takes into account various considerations such as capitalization, hyphens, apostrophes, non-alphabetic characters, and separators to accurately identify the tokens in the text.

In many cases, documents are written using markup languages like HTML or XML, which specify the structure and formatting of the document. Tags are used in these languages to indicate different elements of the document. For example, the tag <h2> is used to indicate a heading element with the text "Overview," and </h2> denotes the end of the heading. The document parser uses the syntax of the markup language to identify the structure of the document based on these tags and other formatting instructions.

The importance of a parser in information retrieval lies in its ability to recognize the structural elements of a document. By understanding the organization and layout of the document, the parser enables more effective indexing and retrieval of relevant information.

The parser helps in extracting important components such as titles, headings, and links, which are crucial for understanding the document's content and context. This enables better organization and categorization of documents in an information retrieval system.

Furthermore, the parser plays a key role in processing user queries. It ensures that the query is properly understood by identifying the syntactic structure and components of the query. This allows for accurate matching of user queries with relevant documents during the retrieval process.

In summary, the parser is an essential component in information retrieval systems as it processes the sequence of text tokens, recognizes structural elements, and enables better organization, indexing, and retrieval of relevant information. Its ability to interpret the structure of documents and queries enhances the overall effectiveness and efficiency of the information retrieval process.

ii) Link Analysis

Link analysis, which involves examining at the links and anchor text on web pages, is an essential part of information retrieval. Information retrieval systems can learn a lot about a web page's popularity and community information by examining these links. PageRank is a well-known method of link analysis that measures a web page's authority and importance based on the quantity and quality of inbound links it receives. The anchor text connected to the links is also taken into account by link analysis because it offers context and additional keywords related to the target page.

In web search applications, the value of link analysis is quite apparent. The relevance and value of web pages are mostly determined by link analysis algorithms used by search engines. Search engines can effectively rank and prioritise search results by taking into account the authority and popularity conveyed through inbound links, providing users with more accurate and dependable information. By adding more information and keywords relevant to the target page, the analysis of anchor text improves search results even more.

However, it is crucial to remember that the importance of link analysis may differ across various information retrieval applications. The impact of link analysis may be lessened in specialised databases or docsument repositories where hyperlinks are less prevalent or meaningful. Nevertheless, in the context of web search, link analysis plays a vital role in improving the accuracy, relevance, and ranking of search results, ultimately enhancing the overall information retrieval experience for users.

5. Menggunakan contoh yang sesuai, terangkan mengapa teknik *pooling* penting dalam penilaian sistem capaian maklumat berdasarkan kaedah koleksi ujian?

*Using a suitable example, explain why pooling technique is important in evaluation of information retrieval systems based on test collection method?*

[3 markah/*marks*]

In the evaluation of information retrieval (IR) systems based on the test collection method, pooling is an essential technique. It plays a crucial role in determining the effectiveness and performance of IR systems.

To give an example, suppose we have a test collection method that consists of a series of queries and a corresponding set of documents for each query. In order to evaluate the retrieval performance of various IR Systems, this test collection is utilised. Following that, each system will provide a ranked list of documents for each query.

The pooling technique starts here. It includes merging the output of different IR Systems and creating a unified set of documents for each query. The pooling process collects all the documents retrieved by different systems, eliminates duplicates, and combines them into a single collection for evaluation purposes.

The relevance judgement is one of the most important factors that contributed to importance in pooling. It is impractical to make exhaustive judgements on every document in a collection. It rapidly grew impossible to source relevance judgements for every topic-document pair in the collection due to the rapid growth in the size of document collections and the high expense of doing so.  The work to produce a complete set of relevance judgements would take more than a researcher's entire lifetime, even for a relatively small test collection with half a million documents and a few tens of topics. Thus, pooling is the technique employed to avoid complete assessment. Pooling method reduces the number of relevance judgements that are necessary in order to accurately assess the effectiveness of an IR system or more importantly establishing the difference between the effectiveness of two systems.

References

Akhoon, I. ul H. (n.d.). *Evaluation of search engines using Advanced Search Technique: A Comparative Analysis of Yahoo and bing*. DigitalCommons@University of Nebraska - Lincoln. https://digitalcommons.unl.edu/libphilprac/2813/

Aliannejadi, M., Zamani, H., Crestani, F., & Croft, W. B. (2019). Asking clarifying questions in open-domain information-seeking conversations. *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval*. https://doi.org/10.1145/3331184.3331265

Azad, H. K., & Deepak, A. (2019). Query expansion techniques for information retrieval: A survey. *Information Processing & Management*, *56*(5), 1698–1735. https://doi.org/10.1016/j.ipm.2019.05.009

Bjelland, J., Canright, G., & Engø-Monsen, K. (1970, January 1). *Link analysis and web search*. SpringerLink. https://link.springer.com/referenceworkentry/10.1007/978-0-387-30440-3_312

Gola, M. (2020a, January 11). *Google vs. Yahoo, which one is best?*. Curvearro. https://www.curvearro.com/blog/google-vs-yahoo-which-one-is-best/

Gola, M. (2020b, January 11). *Google vs. Yahoo, which one is best?*. Curvearro. https://www.curvearro.com/blog/google-vs-yahoo-which-one-is-best/

*Google vs yahoo: Comparing the top two search engines*. Purple Cow Agency. (2020, June 23). https://purplecowagency.com/google-vs-yahoo-comparing-the-top-two-search-engines/

Google. (n.d.-a). *How google's knowledge graph works - knowledge panel help*. Google. https://support.google.com/knowledgepanel/answer/9787176?hl=en#:~:text=The%20Knowledge%20Graph%20allows%20us,it's%20determined%20to%20be%20useful.

Google. (n.d.-b). *In-depth guide to how Google Search works*. Google. https://developers.google.com/search/docs/fundamentals/how-search-works#:~:text=Indexing%3A%20Google%20analyzes%20the%20text,relevant%20to%20the%20user%27s%20query

Gul, S., Ali, S., & Hussain, A. (2020). Retrieval performance of google, yahoo and bing for navigational queries in the field of "Life science and biomedicine." *Data Technologies and Applications*, *54*(2), 133–150. https://doi.org/10.1108/dta-05-2019-0083

Hamid, A. (2017). (PDF) relevance feedback in Information Retrieval Systems - Researchgate. https://www.researchgate.net/publication/320730332_Relevance_Feedback_in_Information_Retrieval_Systems

Lipani, A., Palotti, J., & Lupu, M. (n.d.). Fixed-cost pooling strategies based on IR evaluation measures. https://core.ac.uk/download/pdf/195309754.pdf

Lutkevich, B. (2022, July 7). *What is a parser? definition, types and examples*. App Architecture. https://www.techtarget.com/searchapparchitecture/definition/parser

Saboroff, I. (2021). Overview of TREC 2021. https://trec.nist.gov/pubs/trec30/papers/Overview-2021.pdf

SEJ STAFF Roger Montti June 29, 2022 · 10 min read SEJ ST. (2022, June 29). *Google freshness algorithm: Everything you need to know*. Search Engine Journal. https://www.searchenginejournal.com/google-algorithm-history/freshness-algorithm/

*Understanding query parsers: How search engines process your searches*. Understanding Query Parsers: How Search Engines Process Your Searches. (n.d.). https://marketbrew.ai/understanding-query-parsers-how-search-engines-process-your-searches

Wagenpfeil, S., Kevitt, P. M., & Hemmje, M. (2023). Smart Multimedia Information Retrieval. *Analytics*, *2*(1), 198–224. https://doi.org/10.3390/analytics2010011

*What is meant by Google, yahoo & bing and how to do google, yahoo & bing*. Inventateq. (n.d.). https://www.inventateq.com/about-search-engine-and-how-they-work.php#:~:text=Yahoo%3A%20Yahoo%20Search%20is%20a,title%20must%20contain%20the%20keywords.