Unit 8

Hunting Spiders through Search Engines

FUNCTIONS OF WEB SEARCH ENGINE

Pre-reading Activities

In this unit, you will

- improve your understanding of the target technical words.
- learn about supporting topic sentences with comparison in writing.
- learn how to preview a reading comprehension passage through pre-reading questions to improve comprehension.
- be familiar with the functions of a web search engine.

I. Target Academic Vocabulary

Check out the meanings and functions of the target academic words in a monolingual and bilingual dictionary.

Hierarchical (adj)

Venture (n)

Premise (n)

Crawler (n)

Retrieve (v)

Exclusion (n)

Proximity (n)

Customize (v)

II. Writing development

Supporting topic sentences with comparison

Comparison is defined as "likenesses". A topic sentence can be supported through *similarities* between two things or two aspects of one thing. Some of the common structures of comparison used in a comparative paragraph are introduced.

A. Adjective/preposition (the same/as; similar/to; like)

Examples:

A meta-search engine employs the search results of other search engines **the same** way **as** a searching tool.

A meta-search engine employs the search results of other search engines **similar to** a searching tool.

Like meta-search, a searching tool employs the search results of other search engines.

B. Attached statements (and……too; as so; and …….either; and neither)

Examples:

A meta-search engine employs the search results of other search engines **and** a searching tool does **too**.

A meta-search engine employs the search results of other search engines **and so** does a searching tool.

A meta-search engine does not employ the search results of other search engines **and** a searching tool does not **either**.

A meta-search engine does not employ the search results of other search engines **and neither** does a searching tool.

III. Pre-reading questions:

Read and respond to the questions below, and then discuss them in pair/group.

1. What is a web search engine?

---

2. What are the features of a web search engine?

---

3. Who uses web search engines and why does s/he use them?

IV. Reading comprehension passage

This passage discusses the functions of web search engines focusing on improving a quality search.

FUNCTIONS OF A WEB SEARCH ENGINE

1. Definition and types of web search engines

A web search engine is defined as a software system that is designed to search for information on the World Wide Web. There are three main types of searching tools that have evolved. First, human beings have extensively programmed a system of predefined and hierarchically ordered keywords which is defined as web directory. A web directory does not display lists of web pages based on keywords; instead, it lists web sites by category and subcategory. An example of a well-known general web directory is DMOZ (http://www.dmoz.org). Second, a system generates an "inverted index" by analyzing the texts it locates on the network of a search engine. This type is a commercial venture supported by advertising revenue and thus some of them allow advertisers to have their listings ranked higher in search results for a fee. For example, search engines that do not accept money for their search results make money by running search related ads along with the regular search engine results. The search engines make money every time someone clicks on one of these advertisements. Finally, a meta-search engine is a searching tool that employs the search results of other search engines and aggregates the results into a single list or displays them according to their source. The meta-search engine enables users to enter search criteria once and access several search engines

simultaneously. Meta-search engines operate on the premise that the Web is too large for any one-search engine to index it and more comprehensive search results can be obtained by combining the results from several search engines. This meta-search engine, for instance, may save the user from having to use multiple search engines separately. Examples of meta-search engine are Dogpile (http://dogpile.com) and Yeppi (http://yeppi.com).

The search results are generally presented on lines of results and often referred to as Search Engine Results Pages (SERPs). The information from search results may be used for specialists on web pages. Some search engines also retrieve available data from databases or open directories. Unlike web directories, which are maintained only by human editors, search engines maintain real-time information by running an algorithm on a web crawler. Web search engines work by storing information about many web pages, which are retrieved from the Hypertext Markup Language (HTML) pages. HTML is a standardized annotation system for tagging text files to achieve font, color, graphic, and hyperlink effects on World Wide Web displayed pages. A Web crawler as shown in Figure 1, also known as a spider, automatically follows every link on the site and retrieves the pages.

The site owner can make exclusions by using robot.txt. The contents of each page are then analyzed to determine how it should be indexed (for example, words can be extracted from the titles, page content, headings, or special fields called meta tags). Data about web pages are stored in an index database for use in later queries. A query from a user can be a single or sequence of keywords. The index helps find

information relating to the query as quickly as possible. Some search engines, such as Google, store all or part of the source page (referred to as a Cache) as well as information about the web pages, whereas others, such as AltaVista, store the keywords of pages they find. The cached page always holds the actual search text since it is the one that was actually indexed, so it can be very useful when the content of the current page has been updated and the search terms are no longer in it. Increased search relevance makes these cached pages very useful, but not just because they may contain data that may no longer be available elsewhere.

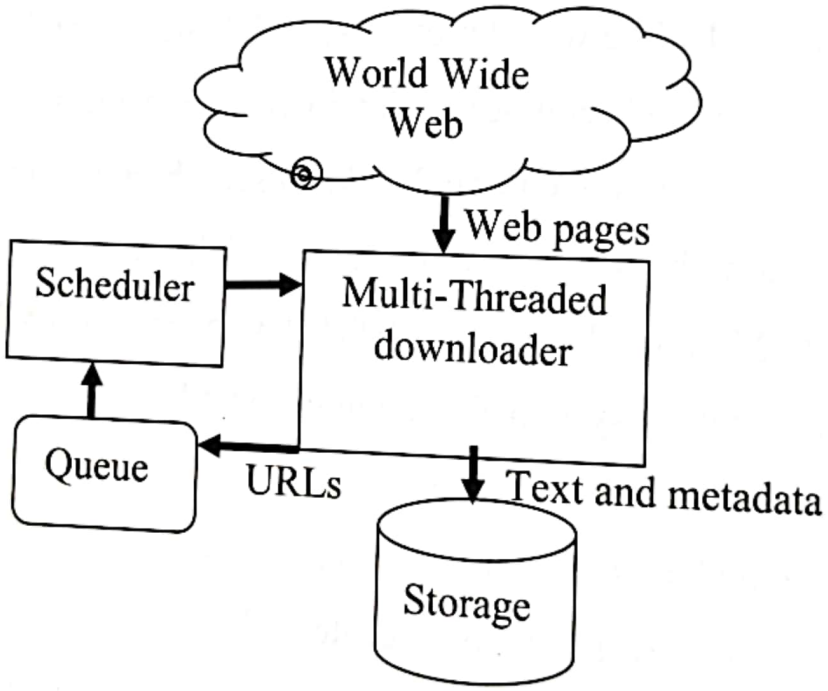

Figure 1: High-level architecture of a standard Web crawler.

When a user enters a query into a search engine (typically by using keywords), the engine examines its index and provides a listing of best-matching web pages according to its criteria, usually with a short summary containing the document title and sometimes parts of the text.

The index is built from the information stored and the method by which the information is indexed. From 2007, Google.com search engine has allowed users to search using dates by clicking on *Tools* tab in the initial search results page, and then selecting the desired date ranges. Most search engines support the use of the Boolean operators AND, OR, and NOT to further specify the search query. Boolean operators are for literal searches that allow the user to refine and extend the terms of the search. The engine looks for the words or phrases exactly as entered. Some search engines provide an advanced feature called proximity search, which allows users to define the distance between keywords in documents. In Google, for example, if you were searching for "Memory AND CPU" before, you could now use "Memory AROUND(1) CPU" instead. This will tell Google that you're looking for Memory and CPU to appear in close proximity to each other. If you want to extend the range a bit, increase the number (e.g. AROUND(2), AROUND(3), etc.). There are also concept-based searches where the search involves using statistical analysis on pages containing the words or phrases you search for. Also, natural language queries allow the user to type a question in the same way one would ask it to a human being. A site with this feature would be ask.com.

The usefulness of a search engine depends on the relevance of the results set it gives back. For instance, there may be millions of web pages that include a particular word or phrase, but some pages may be more relevant, popular, or authoritative than others. Most search engines employ methods to rank the results to list the "best" results first. How a search engine decides which pages are the best matches, and

what order the results should be presented, varies widely from one engine to another. The methods also change over time as Internet usage changes and new techniques evolve.

2. Search engine bias

Although search engines are programmed to rank websites based on their popularity and relevancy, empirical studies indicate various political, economic, and social biases in the information they provide. These biases can be a direct result of economic and commercial processes (e.g., companies that advertise with a search engine can also become more popular in its organic search results), and political processes (such as the removal of search results to comply with local laws).

Biases can also be the result of social processes, as search engine algorithms are frequently designed to exclude non-normative viewpoints in favor of more "popular" results. Indexing algorithms of major search engines skew towards the coverage of US-based sites, rather than websites from non-US-based sites.

3. Customized results and filter bubbles

Many search engines, such as Google and Bing provide customized results based on the user's activity history. This leads to an effect that has been called a filter bubble. The term describes a phenomenon in which websites use algorithms to selectively guess what information a user would like to see, based on information about the user (such as

location, past click behavior and search history). As a result, websites tend to show only information that agrees with the user's past viewpoint, effectively isolating the user in a bubble that tends to exclude contrary information. Prime examples are Google's personalized search results and Facebook's personalized news stream. According to Eli Pariser, who also coined the term Pariser, users get less exposed to conflicting viewpoints and are isolated intellectually in their own informational bubble. Pariser related an example in which one user searched Google for "BP" and got investment news about British Petroleum while another searcher got information about the Deepwater Horizon oil spill and that the two search results pages were "strikingly different". The bubble effect may have negative implications for civic discourse, according to Pariser. Since this problem has been identified, competing search engines have emerged and seek to avoid this problem by not tracking or "bubbling" users.

Post-reading Activities

I. Reading comprehension

Directions: Mark each statement as T (True), F (False), or NG (Not Given) to the information in the reading comprehension passage.

-----1. Three kinds of search engines are discussed in the passage.

-----2. Meta-search engine helps users to have access to several search engines at the same time.

-----3. Some search engines receive data from an unavailable database.

-----4. HTML is a writing style used by some researchers.

-----5. The cached page keeps information as it is normally indexed.

-----6. Search results in a search engine are listed based on the amount of text.

-----7. Search engines are programmed based on only political biases.

-----8. The bubble effect has a positive effect on civic discourse.

Questions ***9-15***: Choose the appropriate letter ***A-C***.

9. A search engine is a software to …….

A. provide information on the internet.

B. design a structure for the information.

C. hunt for information on the internet.

10. Search engines do not make money based on the search results, but ……

A. they make money any time someone receive a search result.

B. advertisers can make money through a search result.

C. they can make money any time someone click on the advertisements.

11. Which one of the following is NOT the feature of the web search engine?

A. Information is stored on the web pages.

B. Information can be stored on the HTML page.

C. Search results are in a line on the SERPs.

12. All of the following statements are TRUE when a user puts a query into a search engine but.....

A. listing of matching web pages is provided.

B. available data may not exist anywhere else.

C. indexes are examined through a web search engine.

13. A helpful search engine relies on the following feature.....

A. the size of a web page.

B. the relevance of the result.

C. the amount of information.

14. Which of the following statement is NOT true based on the reading comprehension passage?

A. Search engine bias is relevant to the economics and commercial processes.

B. A filter bubble is based on the customized results of a users' activity history.

C. Search engines never provide different results for a single query.

15. The meta-search engine provides an opportunity for a user to multiple search engines.....

A. at the same time.

B. one at a time.

C. immediately.

II. Vocabulary activities

Directions: Read each sentence on functions of web search engine stated below. Circle the one word or phrase in parentheses ( ) that has the same meaning as the underlined word in the sentence. Compare your answers with a partner.

1. A system of predefined and hierarchically (*a complete way/on the whole/arrange in order of rank*) ordered keywords that humans have programmed extensively. This type relies much more heavily on the computer itself to do the bulk (*some/little/mass*) of the work.
2. A meta-search engine is a searching tool that employs the search results of other search engines, aggregates (*analyzes/separates/collects*) the results into a single list or displays them according to their source.
3. The information may be used for a specialist on web pages. Some search engines also retrieve (*pull/push/extract*) available data from databases or open directories.
4. Web search engines work by storing information about many web pages, which they retrieve from the Hypertext Markup Language (HTML) page. HTML is a standardized system for tagging (*attaching/joining/hanging*) text files to achieve font, color, graphic, and hyperlink effects on World Wide Web pages.
5. Most search engines support the use of the Boolean operators AND, OR and NOT to further specify (*identify/improve/decrease*) the search query. Boolean operators are for literal searches that allow the user to refine and extend (*indicate/determine/add*) the terms of the search.

6. How a search engine decides which pages are the best matches, and what order the results should be presented, varies widely from one engine to another. The methods also change over time as Internet usage changes and new techniques evolve (*improve/ devalue/ develop*).
7. Indexing algorithms of major search engines skew (*hierarchy/extent/change*) towards coverage of the US - based sites, rather than websites from the non-US-based sites.
8. The bubble effect may have negative implications for civic (*city/old/new*) discourse, according to Pariser.

III. Writing development activities

1. What kind of statements are joined by *and.....too* and *and so*?

---

2. What kind of statements are joined by *and.....either* and *and neither*?

---

3. Write five sentences comparing a laptop with PC using structures of comparison. Get your classmate to provide you with his/her feedback.
