Adaptive Web Search Engine

M. Barouni E. (m.barouni@unb.ca) Ali A. Ghorbani (ghorbani@unb.ca)

Introduction

- **Web Search Engine** (WSE): WSE (e.g. Google) is a program which helps users to find related web pages for a given query.
- **Indexing**: In order to find the related documents (web pages), an indexing mechanism is needed. It consists of two parts:
  - Phrase List: List of all the phrases in the searched space.
  - Phrase Pointer: List of pointers for each phrase in which all of the documents that contains that certain phrase have been indexed.

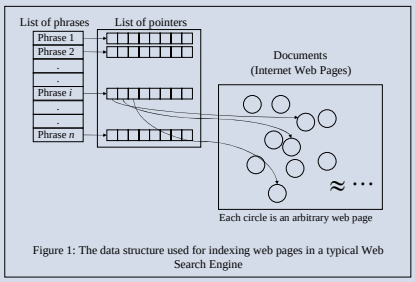

Figure 1: The data structure used for indexing web pages in a typical Web Search Engine

- **Crawler**: Due to dynamic nature of the web, pointer updating (figure 1) is an overwhelming task. Crawlers are hence designed and responsible for browsing the web and updating these indexes.

Phrase Saving Techniques

Before indexing, phrases need to be turned into a canonical format. Some techniques are common to reduce the size of information.

- **Case-folding**: Texts are case insensitive for saving.
- **Stemming**: Saving one instance of the words with the same morphological root.
- **Stopping**: Not saving the words which have no information such as: the, a, it and so on.

Querying Models

- **Boolean querying**: A query consists of some phrases in which AND/OR are located between them. The result of the query is a list of documents (web sites) combined by the intersection/union of the list of pointers for the phrases of the query.
- **Ranked querying**: Search engines calculate a heuristic similarity score that shows the similarity of a query with a document. One easy way can be the coordinate matching method in which scoring of similarity is related to the number of the phrases in the query that appears in the document.

Retrieval Effectiveness

- **Precision**: If there are *l* related documents in the result of a query and the total number of documents in the result is *k*, precision would be *l* / *k*.
- **Recall**: If there are *l* related documents in the result of a query and the total number of related documents in database (e.g. internet) is *N*, recall would be *l* / *N*.

Adaptive Web Search Engine

- **Static WSE**: A WSE is static if the search result of a given query does not change for different users or between time intervals (time interval is the time between updating the index database of a WSE).
- **Adaptive Web Search Engine**: An adaptive WSE changes its results or the order of the list of URLs in the result based on the situation or user's interests.
- **Why Adaptation**: The result of a query is not always suitable for the user. User might not have enough information to fill in the required phrases for querying. The result of the query may contains hundreds unrelated answers. If WSE knows something about user interests, user knowledge, user preferences or the domain of the query, it may find more appropriate answers. An adaptive WSE has the skill to gather some of this information during the life time of system and makes the answers more appropriate to the user's requests.

Approach

- **WSE Trend**: The importance of adaptation in WSE on the one hand, and the complexity of the problem on the other hand have brought about a lot of research in this area. Some approaches focus on adapting based on the document space. The proposed algorithms are supposed to categorize the documents, identify user preferences and retrieve related results. Some other approaches focus on the Query Space. The algorithms try to make an appropriate query for the user, and then send the query to a Search Engine.

  Our attempt is shown in figure 2. The adaptation module performs adaptation in three different parts:
  - It recommends some frequent phrases to the user.
  - It makes a more appropriate query to send to the WSE.
  - It reorganizes the result based on the previous user's selections.

  Figure 3 shows the necessary data structure for Adaptive module.

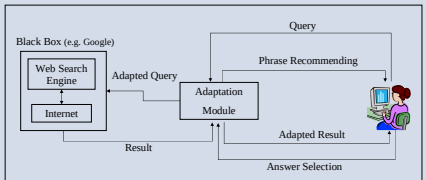

Figure 2: Adaptation Methods in query space.

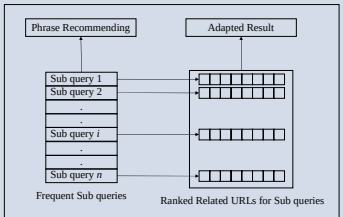

Figure 3: Adaptation Module data structure.