Needs Assessment: Overview

This database project is designed to support research in web search behavior, specifically web search behaviors of "intranet" users inside of a "federal business environment." This database design and relational structure is unique because federal institutions communicate using a hybrid conventional and specialized vocabulary model. Intranet sites in these organizations are ideal to characterize because access to the intranet (internal web page) is "strictly" controlled and can be directly mapped to a single user by their IP address.

While the content on the Intranet site contains information specific to organizational business documents in highly specialized language, it still follows all the same search protocols of a commercial search engine on the "Internet." Institutionally, the intranet, like the Internet, was designed knowing that the capability to create refined, efficient searches was important for both users in receiving efficient and specific results from their web queries.

Our research goal is to characterize an entire years worth of strictly mapped web queries for the purpose of identifying the following data attributes:

1) Identify single user patterns over time (trending).
2) Identify institutional patterns over time.
3) Identify the most commonly requested keyword search terms (frequency of query).
4) Identify persistence of information foraging techniques (link following, indexes, or keyword searches).
5) Characterize mutual information/ co-occurrence (word-pairs, triples, etc.)
6) Determine if keywords form semantic clusters (id possible synonyms)
7) Characterize the sophistication of query or lack thereof and its impact on user satisfaction

Various analytical models will need to be applied in this database. The major inputs to the data model are 400 access and query files. Each pair of files represents a single day of queries. The access file identifies the individual computer on the intranet from which the web information was accessed. The query file contains corresponding keyword, link, or index queries. The query file contains specific keyword, link, index query information, along with date and time. The file information can be correlated by time.

Access_file (IP address, date, time, Query/Command(CSS)/html index call)
Query_file (date, time, Query/index/keyword search)

Business Rules & Mini-World Description

One of the business rules that must be applied to this data is that the IP address must be anonomyzed. A unique identifier will have to be assigned to each unique IP address. In order to track user trends over time, we will need to map the unique id for IP's across all 400 days. It will be important to characterize semantic concepts formed by words, phrases, and n-grams.

Another business rule applies to the instances of multiple words in a query; "Ultraseek" is instructed to match words consecutive order. This means (order) must be associated with each word in a query; a change in order of query words constitutes a "unique" query.

Next, to better illustrate the number of entities in our database design project, please see Figure 1 below.

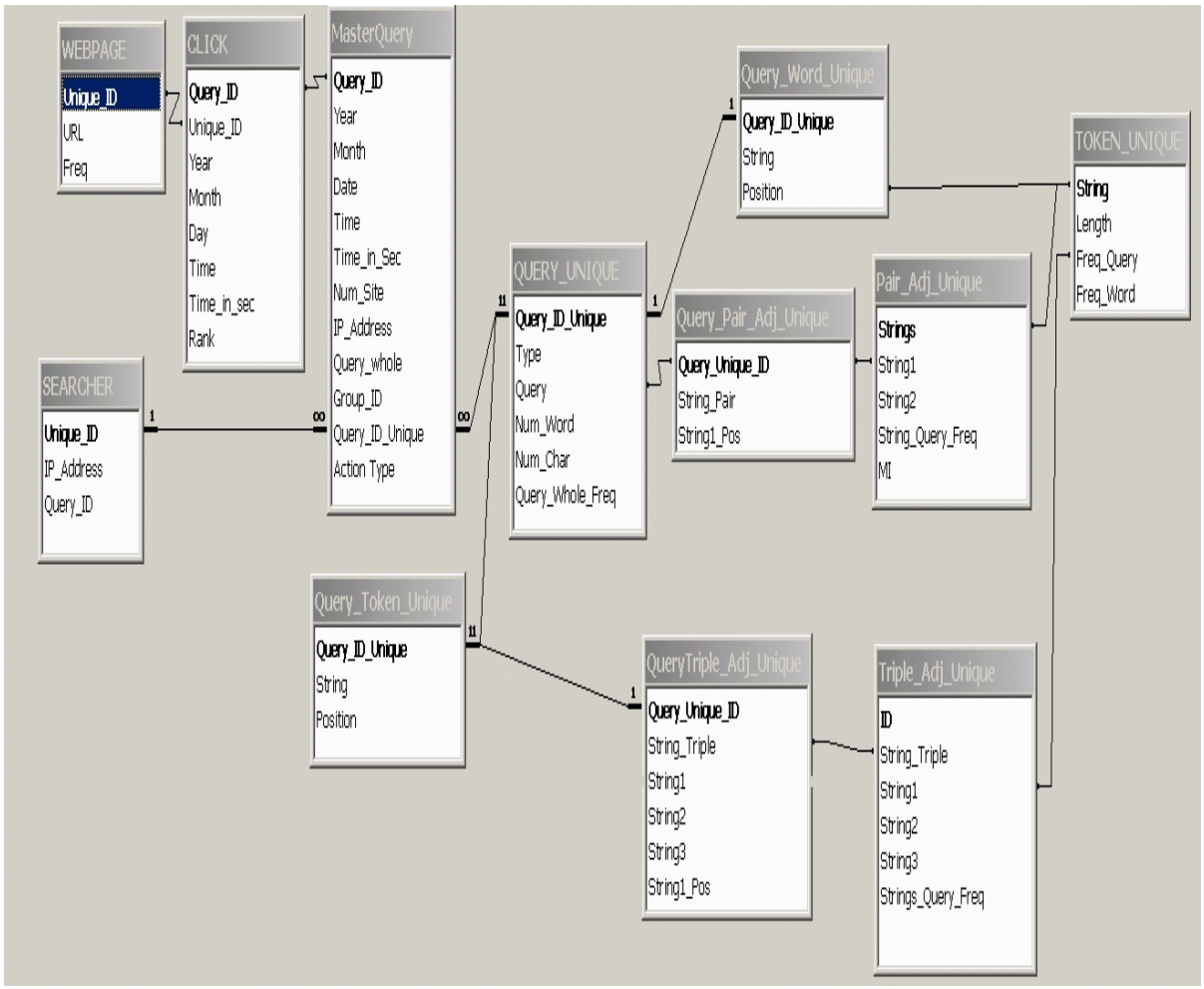

FIGURE 1

There are 12 entities in our database design project. Each entity is uniquely identified with other entities in the database design.

1. The first entity, **WEBPAGE**, has a "one-to-one relationship" with the second entity, **CLICK**.

2. The next entity, **CLICK**, has a "one-to-one relationship" with the **WEBPAGE** entity, and a "one-to-one relationship" with the **MasterQuery** entity. In both of these relationships, the attribute "Query_ID" is the same in all three entities.

3. As stated, the **MasterQuery** entity has a "one-to-one relationship" with the **CLICK** entity. In this relationship, the attribute "Query_ID" is the common attribute. The **MasterQuery** entity has a "many-to-one" relationship with both the **SEARCHER** entity and the **QUERY_UNIQUE** entity. In both relationships the "Unique_ID" attribute is shared amongst the relationships.

4. The **SEARCHER** entity has only the "one-to-many relationship" with the **MasterQuery** entity that was previously mentioned, where the "Unique_ID" attribute is the common attribute.

5. The **QUERY_UNIQUE** entity has four relationships:
   1) A "one-to-many" relationship with the **MasterQuery** entity, with the "Query_ID_Unique" being the common attribute.
   2) A "one-to-one" relationship with the **Query_Token_Unique** entity, with the "Query_ID_Unique" being the common attribute.
   3) A "one-to-one" relationship with the **Query_Pair_Adj_Unique** entity, with the "Query_ID_Unique" being the common attribute.
   4) A "one-to-one" relationship with the **Query_Word_Unique** entity, with the "Query_ID_Unique" being the common attribute.

6. The **Query_Token_Unique** entity has two relationships. First, it has a "one-to-one relationship" with the **QUERY_UNIQUE** entity, with "Query_ID_Unique" being the common attribute. And, second, it has a "one-to-one relationship with the **QueryTriple_Adj_Unique** entity, with "Query_ID_Unique" being the common attribute.

7. As mentioned, the **Query_Word_Unique** entity has a "one-to-one relationship" with the **QUERY_UNIQUE** entity. However, the **Query_Word_Unique** entity also has a "one-to-one relationship with the **TOKEN_UNIQUE** entity, with "String" being the common attribute.

8. The **Query_Pair_Adj_Unique** entity has a "one-to-one relationship" with the **Pair_Adj_Unique** entity, with the common attribute being "String 1." The **Query_Pair_Adj_Unique** entity also has a "one-to-one relationship" with the **QUERY_UNIQUE** entity, with the common attribute being "**Query_ID_Unique."**

9. As stated previously, the **QueryTriple_Adj_Unique** entity has a "one-to-one relationship" with the **Query_Token_Unique** entity, with "Query_ID_Unique" being the common attribute. However, the **QueryTriple_Adj_Unique** entity also has a "one-to-many relationship" with the **Triple_Adj_Unique** entity, with the attributes being: "String_Triple," "String1," "String2," and "String3."

10. The **Pair_Adj_Unique** entity has the "one-to-one relationship" mentioned previously with the **Query_Pair_Adj_Unique** entity, with "Strings" being the common attribute; however, via intersection, it also has a "one-to-one relationship" relationship with the entity **Token_Unique**. The attribute in this relationship is "strings" also.

11. The **TOKEN_UNIQUE** entity has the "one-to-one relationship" with the entity **Pair_Adj_Unique**, via intersection, previously mentioned. The **TOKEN_UNIQUE** entity also has a second "one-to-one relationship" with the entity **Query_Word_Unique**, which was previously mentioned, where "string" is the common attribute. This **TOKEN_UNIQUE** entity has a third "one-to-one relationship" with the entity **Triple_Adj_Unique**, in which the common attribute is "Freq_Query."

12. **Triple_Adj_Unique** is the twelfth entity and, as mentioned, has a "one-to-one relationship" with the **TOKEN_UNIQUE** entity, in which "Freq_Query" is the common

attribute. **Triple_Adj_Unique** also has a second relationship, a "many-to-many" relationship with the **QueryTriple_Adj_Unique** entity, where the attributes include: "String_Triple," "String1," "String2," and "String3."

Please see Figure 2 below, in order to better illustrate the 12 entities, the types of entities, their identifier, the attributes, and the domain of each entity.

Entity Table

ID1	ID	ENTITY	TYPE	IDENTIFIER	ATTRIBUTES	DOMAIN
11	11	Triple_Adj_Unique	WEAK	ID	ID String_Triple String1 String2 String3 Strings_Query_Freq	Inclusive
2	2	CLICK	WEAK	Query_ID	Query_ID Day Unique_ID Time Year Rank Month Time_in_sec	Inclusive
3	3	MasterQuery	REGULAR	Query_ID	Query_ID Year Month Date Time Time_in_sec Num_Site IP_Address Query_whole Group_ID Query_ID_Unique Action Type	Inclusive
9	9	Query_Word_Unique	WEAK	Query_ID_Unique	Query_ID_Unique String Position	Inclusive
6	6	Query_Token_Unique	WEAK	Query_ID_Unique	Query_ID_Unique String Position	Inclusive
5	5	QUERY_UNIQUE	WEAK	Query_ID_Unique	Query_ID_Unique Type Query Num_Word Num_Char Query_Whole_Freq	Inclusive
7	7	QueryTriple_Adj_Unique	WEAK	Query_Unique_ID	Query_Unique_ID String_Triple String1 String2 String3 String1_Pos	Inclusive
8	8	Query_Pair_Adj_Unique	REGULAR	Query_Unique_ID	Query_Unique_ID String_Pair String1_Pos	Inclusive
12	12	TOKEN_UNIQUE	REGULAR	String	String Length Freq_Query Freq_Word	Inclusive
10	10	Pair_Adj_Unique	WEAK	Strings	Strings String1 String2 String_Query_Freq MI	Inclusive
4	4	SEARCHER	REGULAR	Query_ID	Unique_ID IP_Address Query_ID	Exclusive
1	1	WEBPAGE	WEAK	Unique_ID	Unique_ID URL Freq	Exclusive

FIGURE 2

There are 12 relationships that are formed from these 12 entities from this module. For a better description of these relationships, their definitions, degrees, cardinalities, and attributes, please see figure 3 below.

Relationship Table

ID	RELATIONSHIP	DEFINITION	DEGREE	ATTRIBUTES	CARDINALITY
2	CLICK	WEAK	Query_ID	Query_ID Day Unique_ID Time Year Rank Month Time_in_sec	Inclusive
4	CLICK:MasterQuery	STRONG	TERNARY	Year Month Time Time_in_sec	M:N
12	MasterQuery:QUERY_ UNIQUE	WEAK	UNARY	Query_ID	1:1
6	Pair_Adj_Unique:TOKEN_ UNIQUE	WEAK	UNARY	Query_ID_Unique String Position	Inclusive
8	Query_Token_ Unique: QUERY_UNIQUE	WEAK	UNARY	Query_Unique_ID	1:1
7	QUERY_UNIQUE:Query_ Word_Unique	WEAK	UNARY	Query_ID_Unique	1:1
9	Query_Word_Unique	WEAK	Query_ID_Unique	Query_ID_Unique String Position	Inclusive
5	Query_Word_Unique: TOKEN_UNIQUE	WEAK	UNARY	String	1:1
3	QueryTriple_Adj_Unique: Triple_Adj_Unique	STRONG	STRONG	String_Triple String1 String2 String3	M:N
10	SEARCHER:MasterQuery	STRONG	BINARY	Query_ID IP_Address	M:N
11	Triple_Adj_Unique	WEAK	ID	ID String_Triple String1 String2 String3 Strings_Query_Freq	Inclusive
1	WEBPAGE:CLICK	WEAK	UNARY	Unique_ID	1:1

FIGURE 3