2009

University of North Carolina Wilmington

Master of Science in

Computer Science and Information Systems

Proceedings

https://csbapp.uncw.edu/mscsis

EVALUATION OF SPELLING CORRECTION AND CONCEPT-BASED SEARCHING MODELS IN A DATA ENTRY APPLICATION

Royce Anthony Nobles

A Thesis Submitted to the
University of North Carolina Wilmington in Partial Fulfillment
of the Requirements for the Degree of
Master of Science

Department of Computer Science
Department of Information Systems and Operations Management
University of North Carolina Wilmington

2009

Approved by

Advisory Committee

_____ Dr. Thomas Janicki _____ _____ Dr. Darwin Dennison ____

_____ Dr. Curry Guinn ____
Chair

Accepted By

Dean, Graduate School

TABLE OF CONTENTS

ABSTRACT

Much advancement has taken place in the field of Information Retrieval toward improving the efficiency and ease with which documents can be located. While research has improved the performance of text searches on massive, domain independent databases (e.g., Google™), most domain-specific applications available to consumers use more limited search techniques. In particular, nutrition software applications for logging and analyzing an individual's daily food consumption commonly employ simple Boolean text matching to locate food items from within a database, often resulting in long search times and difficulty locating foods.

DINE Healthy is a nutrition assessment and diet improvement software program developed by DINE Systems, Inc. and widely used in college and university nutrition classes. The current version (7.0) accomplishes food searching by matching user query strings with food item descriptions using the common Boolean Search approach. This research evaluates the effectiveness of enhancing the DINE Healthy food search system with advanced Information Retrieval techniques including spelling error correction and concept-based searching.

To facilitate independent evaluation of these techniques, two additional food search versions are developed for use with the DINE Healthy software. The first extends the current system with only spelling error correction while the second includes both spelling error correction and concept-based searching. Analysis of usage data from a population of university students using each food search version demonstrates that improvements in speed of data entry, a reduction in the rate of failed searches, and a decreased number of queries required to locate food items can be achieved using the enhanced search versions.

ACKNOWLEDGEMENTS

So many kind and generous people have contributed to this thesis and to my graduate education as a whole. I would like to thank Dr. Curry Guinn whose leadership, vision and incredible talent guided me throughout this research. His remarkable insight and genuine enthusiasm kept me constantly motivated to learn. I am very grateful to Dr. Tom Janicki for being a mentor and for teaching me to always see the big picture. Through our conversations I have come to a deeper understanding of my profession. I extend my sincerest thanks to Drs. Darwin and Kathryn Dennison for tirelessly giving their support, kindness, and wisdom. Knowing and working with them has shaped my education and my life.

I am very grateful to Dr. Gene Tagliarini for teaching me to be a scientist and to Dr. Doug Kline for encouraging me to develop expertise in whatever I do. My thanks to Dr. Ron Vetter for his incredible leadership and to the faculty and staff for all they do to make the MSCSIS program an outstanding academic experience.

Finally, I am most grateful to my family and friends for the support, encouragement and motivation to see this through. In particular, I would like to thank my wife Karla for her understanding, patience, and thoughtful advice. I could not have done it without her.

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1: Introduction

The field of Information Retrieval has been the subject of much research and advancement in recent years. Performance gains in text searching of large, domain independent databases (e.g. Google™) have resulted in fundamental changes to the way people access data. However, most domain-specific applications available to consumers continue to use more limited search techniques. In particular, nutrition software applications for logging and analyzing daily food intake commonly use simple Boolean text matching to locate food items from within a database, often resulting in long search times and difficulty recording food consumption.

DINE Healthy is a nutrition assessment and diet improvement software program developed by DINE Systems, Inc. and used in college and university nutrition classes[1]. Students use the software to create daily food intake records by searching for and recording the foods they consume. Food intake records can be analyzed for 122 different nutrients and food components, and various analysis reports and graphs are available.

The software contains a database of more than 10,000 unique food items, each having a text description up to 255 characters in length. These descriptions contain a series of comma separated phrases describing each food (e.g. name, preparation, and manufacturer). The practice of grouping this information into a single field stems from the fact that many food items are obtained from the USDA Nutrient Database for Standard Reference, which uses this format.

1.1 Overview

This research evaluates the effectiveness of enhancing the DINE Healthy food search with advanced Information Retrieval techniques including spelling error correction and concept-based searching and sorting. To facilitate evaluation of these enhanced searching techniques, the DINE

[1] DINE represents the phrase Diet Improvement and Nutrition Education.

Healthy food search is analyzed and used as a basis for comparison. Two additional food search versions are created for use with the DINE Healthy software to examine the performance contributions of spelling error correction and concept-based searching independently.

A sample population of UNCW students is divided into three groups consisting of a control group using the current DINE Healthy food search, an experimental group using the enhanced food search with spelling error correction, and a second experimental group using the enhanced food search with both spelling error correction and concept-based searching. Each subject is provided with a packet of food item pictures and asked to locate the pictured food items using the DINE Healthy software. Usage data for each of the three food search versions are collected from subjects in the form of XML search log entries.

The usage data collected from each group is analyzed to determine mean search time, percentage of searches that fail to return a valid result, total number of sub-queries required, and mean number of irrelevant results per query. Analysis demonstrates that improvements in speed of data entry, a reduction in the rate of failed searches, and a decreased number of queries required for locating food items can be achieved using the enhanced search versions. These results show that improvements in recall and search time can be achieved without significantly negative impact on precision.

1.2 The DINE Healthy 7 Food Search

Nutrition assessment software programs are often used successfully for nutrition education research and practice, and provide health professionals and researches with a number of advantages over older, manual methods of analysis. However, users of these programs have historically faced limitations concerning speed of data entry and difficulty locating foods resulting from spelling and typographical errors (Probst and Tapsell 2005). This may be

explained in part, because the food search systems of these applications typically lack advanced searching techniques commonly found in large, domain-independent applications.

As an implementation of the common Boolean Search model (see Section 3.2), the DINE Healthy food search is an ideal candidate for experimentation along these lines. Current enhancements for filtering results by category are easily integrated with any text-based search, and the addition of advanced searching techniques can be done without noticeable interface modification. This ensures that all subjects are presented with a consistent user interface.

DINE Healthy food searches are initiated by entering queries into a textbox (on the Food Record, Recipe Record, and Food Explorer windows) and clicking the "Search" button, or pressing enter. This action displays the Food Search Results window, as shown in Figure 1.1 after entering the query "peanut butter and jelly". A textbox and search button for entering sub-queries is located near the upper left of the window, and the search results grid and a list for filtering search results by category occupy the central portion.

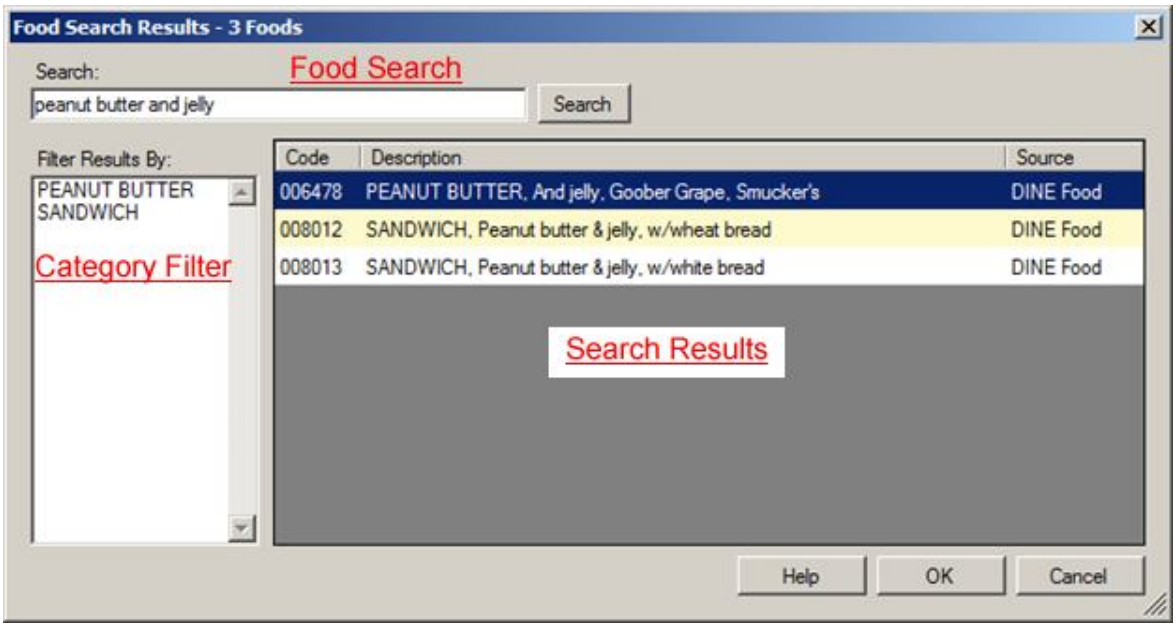

Figure 1.1 DINE Healthy Food Search Results Window

When a user enters a query, it is parsed into words (or terms) which are compared to text descriptions of all foods in the DINE Healthy food database. The set of foods with descriptions containing ALL search terms (an implicit AND operation) are displayed in the search results grid. Assuming that some food descriptions match the query and are returned as results, filter items known as categories are automatically generated in the category list.

Categories are generated by selecting and grouping all text left of the first comma in each food description matching a query string. Selecting a category from the category filter list limits the search results to only food items with descriptions that begin with the selected category. Figure 1.2 illustrates the effect of selecting the "SANDWICH" category filter for a search on "peanut butter and jelly".

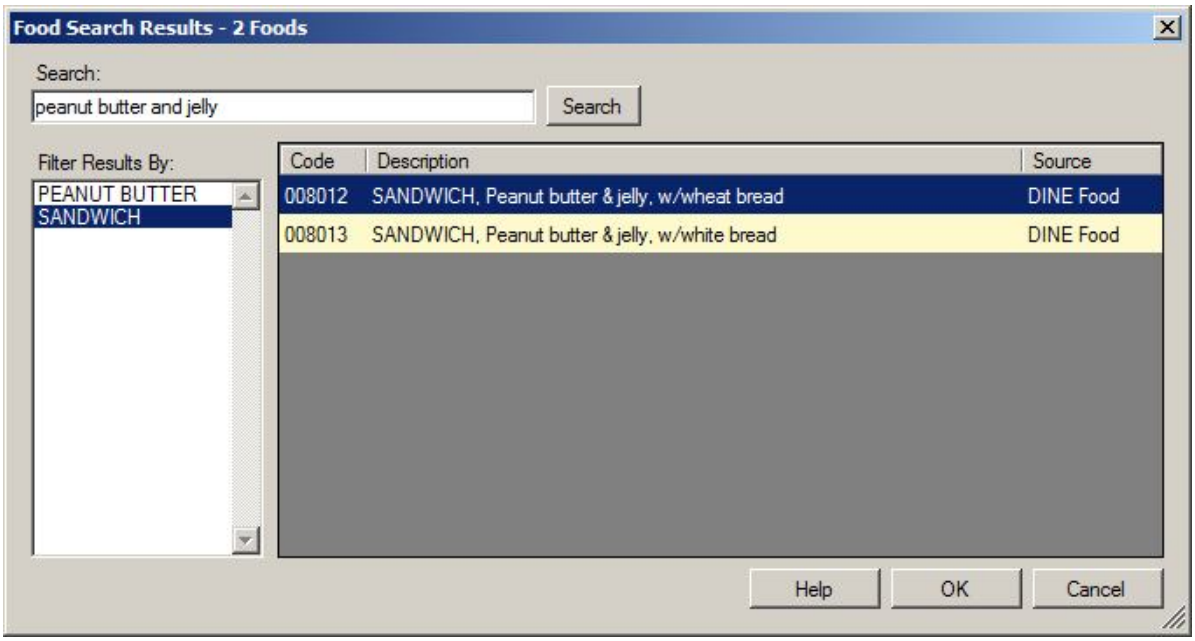

Figure 1.2 Filtered DINE Healthy Food Search Results Window

Users can also refine the search by entering another query into the textbox and clicking the "Search" button or pressing enter. Entering a new query from the Food Search Results window

resets the category filter removing any previous selections. Sub-queries and category filter selections can be repeated any number of times during the course of a single food search.

1.3 Examples of Known Limitations

The enhanced search models are designed to improve upon several known limitations of the DINE Healthy food search, which are:

- It lacks the capacity to recognize and correct spelling errors, thus misspelled words in text queries often prevent the display of search results. Figure 1.3 shows the result of searching on the misspelled word “brocoli” (should be “broccoli”). The correctly spelled query returns 66 food items, while the query containing the misspelling returns none.

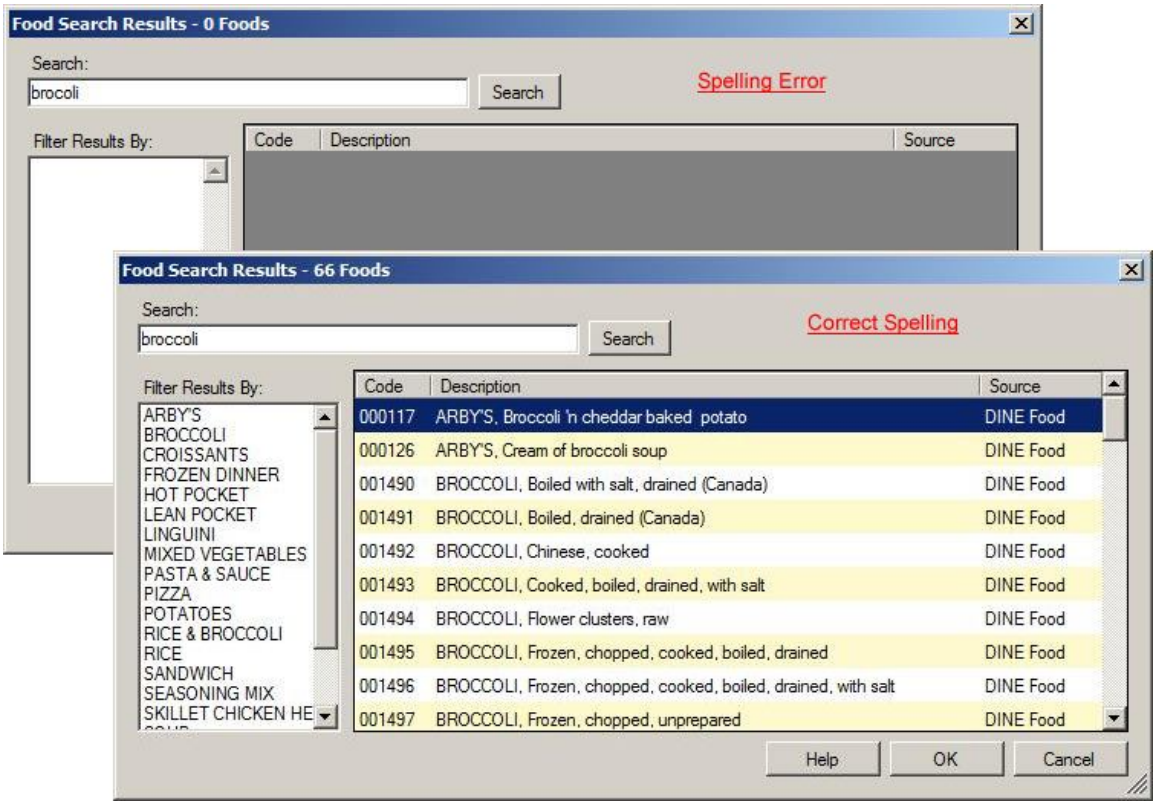

Figure 1.3 Food Search Results with Spelling Error

- It uses simple Boolean text matching, and does not generalize to concepts though several accepted names for individual food items often exist (i.e. soda, pop, cola, soft drink). Figure 1.4 shows the results generated by the current DINE Healthy food search for each of the following queries: "grape soda" (1 result), "grape pop" (2 results), and "grape drink" (17 results).

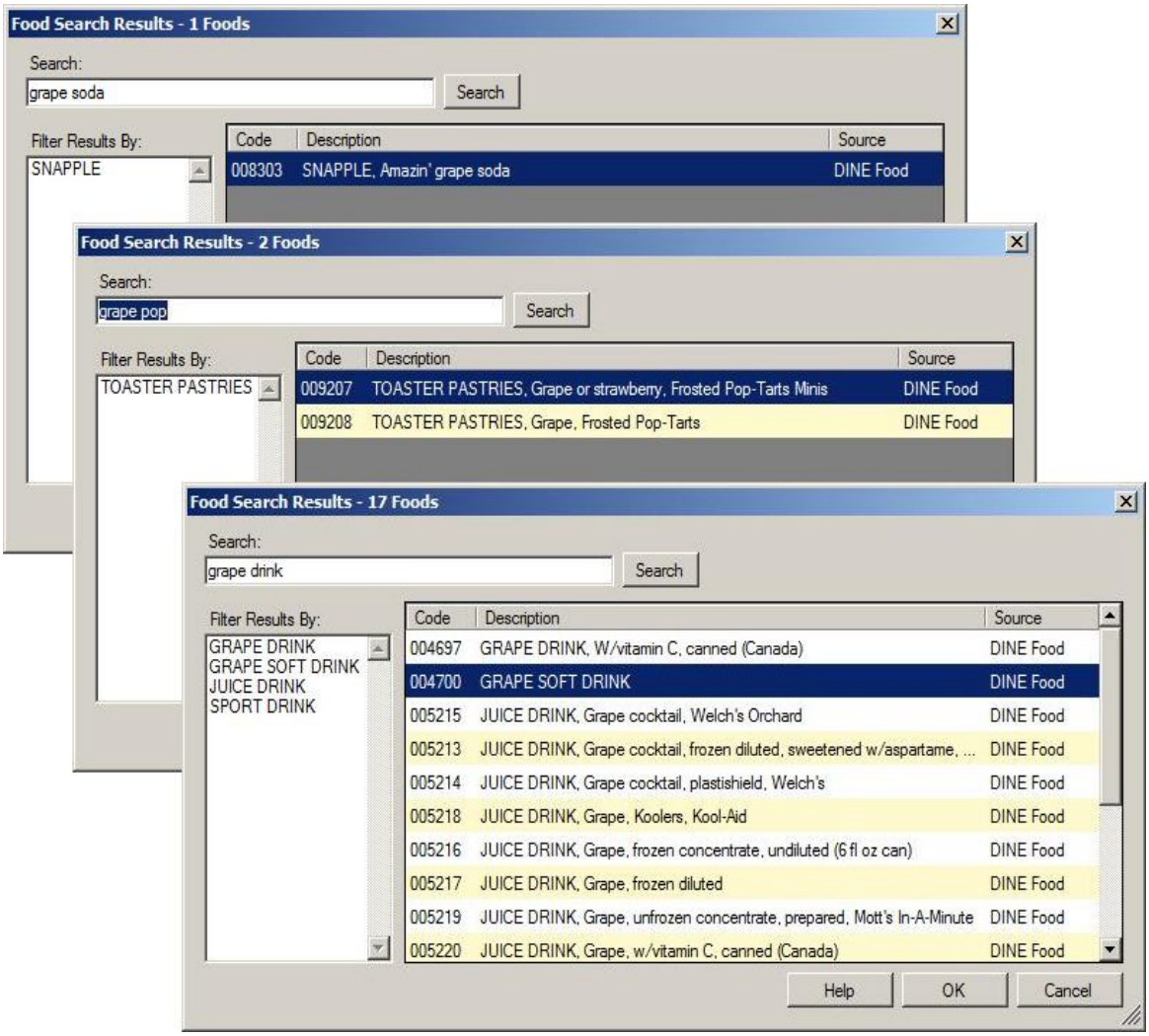

Figure 1.4 Result of Food Searches on Synonyms

- DINE Healthy food search results are sorted alphabetically by food item description. This often results in the desired food item being positioned below numerous unrelated

foods in the results grid. Figure 1.5 shows the result of a search on the string "water". Note that although "water" is the subject of the search, it occurs near the bottom of a list of 89 food items.

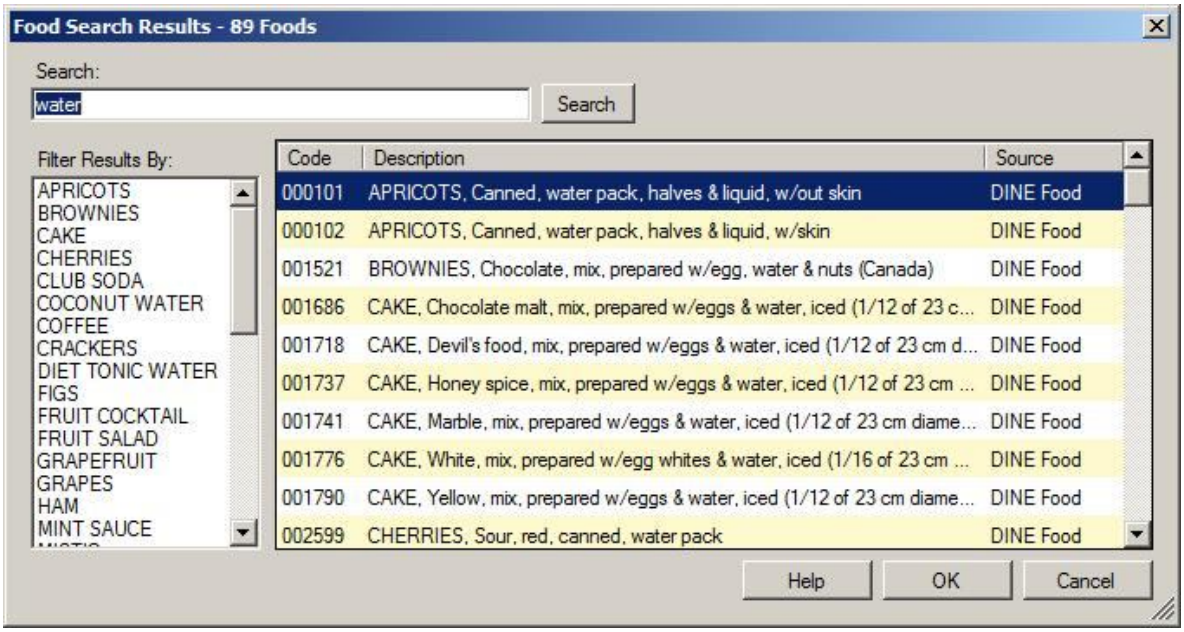

Figure 1.5 Result of Food Search for Water

1.4 Hypotheses

Information Retrieval systems are typically evaluated in terms of recall and precision, and as DINE Healthy is a data logging application, data entry time is also considered. To evaluate the effectiveness of the enhanced food search models with human subjects, the following four hypotheses are tested:

- H_A 1: Reduction of Mean Search Time

"Subjects using the enhanced food search models will take less time to complete the assigned food searches than subjects using the current DINE Healthy food search."

- H_A 2: Reduction of Failed Searches

"*Subjects using the enhanced search models will initiate fewer searches that fail to return an acceptable selection than subjects using the current DINE Healthy food search.*"

- H_A 3: No Significant Negative Impact on Precision

"*Subjects using the enhanced search models will not receive significantly more search results than subjects using the current DINE Healthy food search.*"

- H_A 4: Reduction of Search Refinements

"*Subjects using the enhanced search models will require fewer search refinements to locate food items than subjects using the current DINE Healthy food search.*"

CHAPTER 2: Review of Food Search Systems

Nutrition assessment software programs vary greatly in terms of features and design, but each offers some mechanism for logging and analyzing the nutritional content of an individual's daily food intake. Nutritionists, dieticians and other health professionals use these programs to assist their clients with weight maintenance and the management of nutritionally affected diseases such as diabetes, renal failure, and certain types of cancer. Athletes and fitness-conscious individuals use such software to increase physical fitness and endurance, maintain proper bodyweight, and improve health by regulating the intake of specific nutrients. These programs are also commonly used in university nutrition education classes to teach nutrition awareness and train future health professionals.

Research exists comparing nutrition assessment software programs based on numerous criteria (Probst and Tapsell 2005), but surprisingly little information is available specifically comparing the food search systems of these applications. Some, such as the Healthy Eating Index (HEI 2008), are easily identified as using Boolean Search methods, but others are less obvious. The food search systems of five nutrition assessment software programs are reviewed here to provide perspective.

2.1 Nutribase EZ Edition 7.10

Nutribase EZ Edition 7.10 is a diet, nutrition and fitness improvement software program developed by CyberSoft of Phoenix, Arizona. Individuals seeking assistance with personal weight loss and dietary improvement compose the primary market for this software, and it is designed for novice-level users. The food search interface is integrated into the upper left region of the Food Log window as depicted in Figure 2.1.

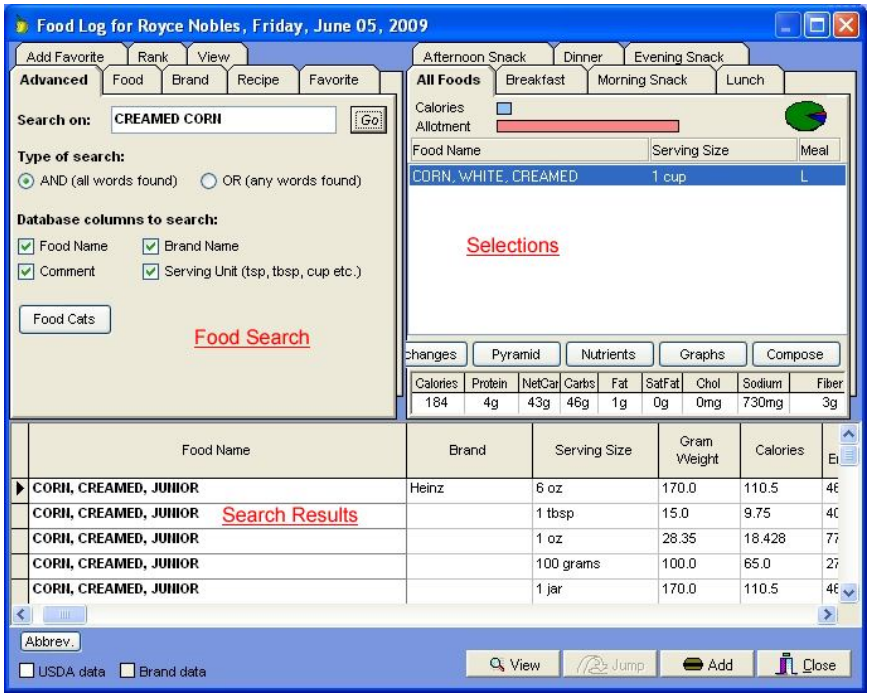

Figure 2.1 Nutribase EZ Edition 7.10 Food Log Window

The list of selected food items is located on the right of window, and the food search results grid occupies the lower portion. To begin searching foods, the user enters a query (i.e. “creamed corn”) into the field labeled “Search on:” and clicks the “Go” button. The query text is parsed into terms and compared with text descriptions of each food item. Search results are displayed unordered, with no apparent alphabetical ordering or relevance ranking.

The user may select the desired implicit operator for linking search keywords by means of two option buttons labeled “AND (all words found)” and “OR (any words found)”. Multiple columns, such as comment, brand, and serving unit can also be searched in addition to the food description. This is accomplished by selecting from a group of four checkboxes located beneath the implicit operator option buttons.

The search system does not apply any sort of spelling error detection or correction strategies, and misspelled search terms typically result in a message box warning that no results matching

the search criteria were found. Concept-based searching and sorting techniques do not appear to be used and are not mentioned in documentation.

2.2 DietPower 4.4

DietPower 4.4 is a weight loss and nutrition coaching software program produced by Diet Power, Inc. of Danbury, Connecticut. Though marketed primarily to individuals seeking to lose weight, this software is designed for use by the general public and with nutrition education courses. The food search interface is integrated within the main Food Log window and contains the search results grid and selected food items grid, as shown in Figure 2.2.

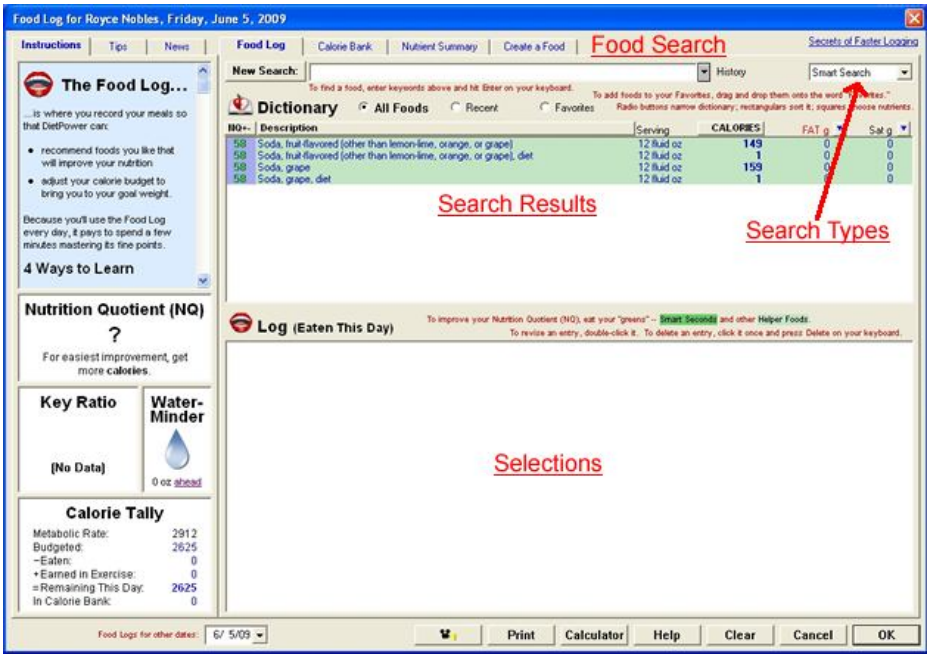

Figure 2.2 DietPower 4.4 Food Log Window

Rather than a single food search system, the interface includes a dropdown for selecting one of four search types (Category Search, Incremental Search, Keyword Search, and Smart Search). The category search simply allows the user to select one of 72 predefined categories for which all food items are displayed.

When using the Incremental Search, an alphabetical list of search results is generated by matching the leftmost characters in the food description with the query string. For example, typing “w” displays all food items having descriptions beginning with “w”, typing “wa” displays all food items having descriptions beginning with “wa”, and so on.

The Keyword Search accepts a query string composed of up to ten individual keywords separated by spaces and/or the OR operator. If terms are separated by a space, the AND operator is used implicitly. This search includes automatic spelling error correction and results are sorted alphabetically by food description. Concept-based searching is not available, although synonyms are reported to be included and searched for some foods.

The Smart Search is an extension of the Keyword Search that attempts to display the most likely results in blue at the top of the result list. This appears to be accomplished by comparing the left-most term in each food item description with the query text and presenting any matching food items first in the alphabetically sorted result list.

The DietPower 4.4 food search possesses a large array of search options and capabilities compared to other nutrition assessment software programs, and the online documentation includes a detailed explanation of these features. However, some search related issues can arise while using the system. For example, when conducting a Keyword or Smart Search, it is necessary to click the “New Search” button to prevent the system from searching only the results of a previous search. This can result in the user receiving a message box stating that no matches were found when the query string would otherwise produce results.

2.3 Food Processor SQL 10.2.2

Food Processor is a diet analysis and health planning software program developed by ESHA Research of Salem, Oregon for more than 25 years. ESHA software is primarily used by

nutritionists, dieticians, and universities, and a student version is distributed with certain nutrition textbooks. The food search interface used in the Food Processor SQL 10.2.2 program consists of a search area occupying the upper portion of the window and a result grid composing the lower portion as shown in Figure 2.3. Food item selections are recorded on a separate window, similar to the DINE Healthy software.

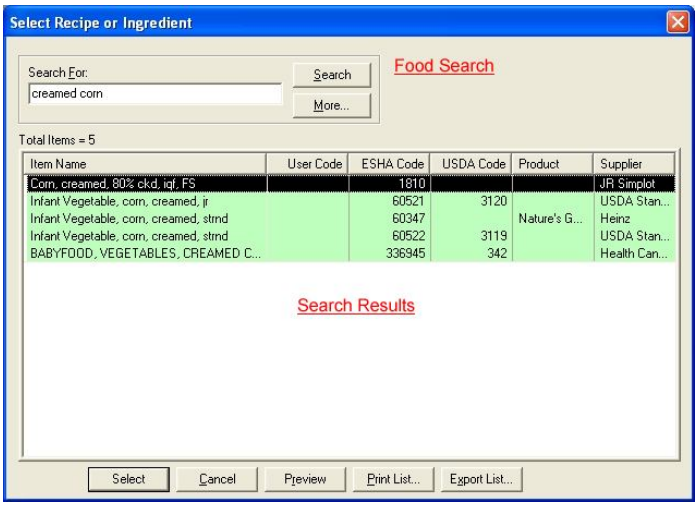

Figure 2.3 Food Processor SQL 10.2.2 Food Search Window

When a search is performed, the query text is parsed into terms which are joined with an implicit AND operator. The query text remains in the "Search For" field for use with sub-queries, and any matching results are displayed in the results grid. Search results are not sorted alphabetically, and no apparent relevance ranking scheme is used. Clicking the "More" button opens a window with several options including filtering by item name, USDA code, ESHA code, user code and group.

The Food Processor SQL 10.2.2 food search system does not make use of any spelling error correction or concept-based searching techniques. Most searches are somewhat slow compared

to other programs reviewed (typically several seconds are required to return results), and only 400 results can be returned per query.

2.4 BeNutrifit 1.7

BeNutrifit 1.7 is a nutrition, diet and fitness software program developed by CMC Enterprises of Scottsdale, Arizona. The software is typically used by bodybuilders, athletes, and individuals seeking to manage bodyweight and increase personal fitness. The food search interface is located on the upper left of the main application window as shown in Figure 2.4. Search results are displayed in a grid positioned directly below the food search interface, and food item selections are displayed in a grid below the search results.

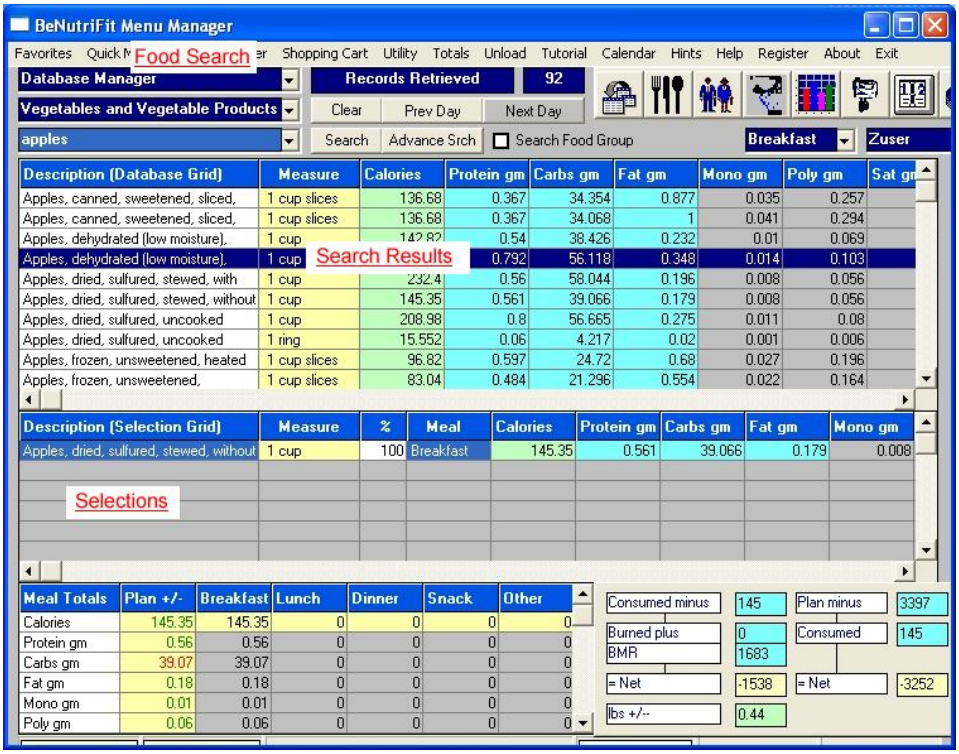

Figure 2.4 BeNutrifit 1.7 Software

Food searches are accomplished by first making a selection from a database dropdown, then selecting from a food category dropdown, and finally typing a query into the search text box and

clicking “Search” (pressing enter while the cursor is positioned in the search text box will not initiate a search).

It is important to note that improper selection of food category will prevent the system from correctly locating food items (I.e. searching for “apples” while the “Poultry Products” category is selected returns no results). Additionally, there is no selection allowing for a search over all categories, so care must be taken to correctly identify the category for each food item.

Query strings are not tokenized or processed with implicit operators such that some part of a food description must match the query string exactly in order for a match to occur (i.e. a query string of “Dried Apples” does not return the food item “Apples, Dried”). The BeNutrifit 1.7 food search system does not make use of spelling error correction or concept-based searching, and results are displayed sorted alphabetically.

2.5 NutriGenie Optimal Nutrition 7.5

NutriGenie Optimal Nutrition 7.5 is a nutrition software program produced by NutriGenie of Stanford, California. The primary market for this software consists of individuals seeking to improve health through good nutrition. The food search interface is integrated into the main application window directly above the search results list as shown in Figure 2.5. This system is unique among those reviewed in that the food item selection list is located on the left side of the window opposite the food search interface and search results list.

By default, food searches are initiated by selecting one of 30 food categories from a dropdown to display all food items belonging to that category. In order to perform a text-based search, the user must click a button on the main toolbar with an image depicting a magnifying glass. This action replaces the category dropdown located above the search results list with a textbox for entering query strings. Once a query string has been entered, the user must click a

blue button located to the right of the textbox (pressing enter with the curser located in the textbox will not initiate a search).

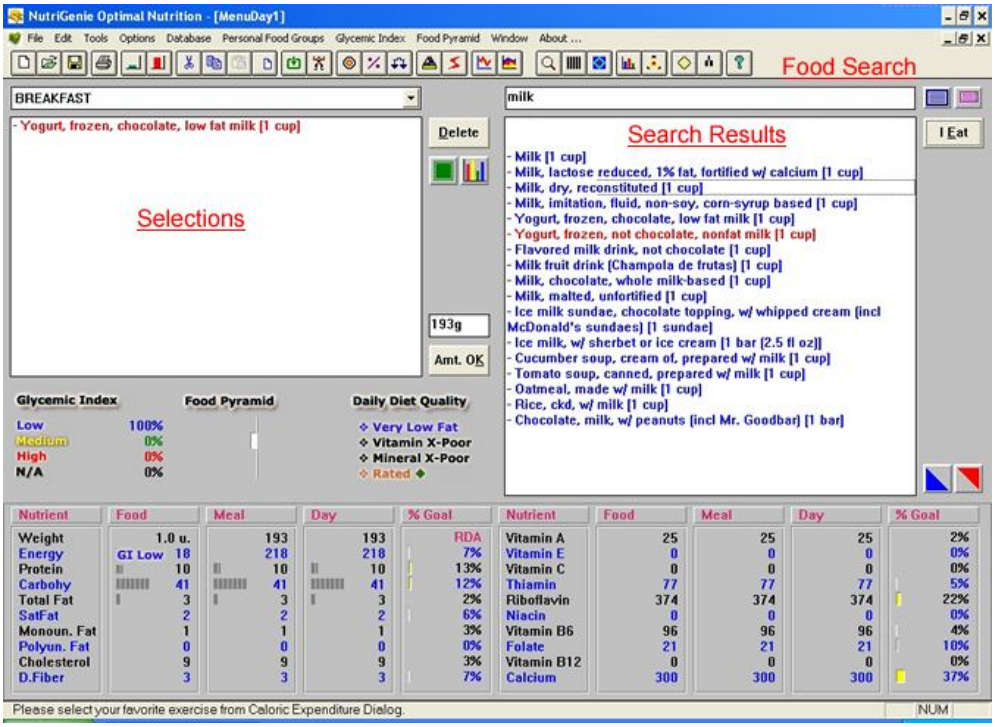

Figure 2.5 NutriGenie Optimal Nutrition 7.5 Software

As with the BeNutrifit 1.7 search system, individual terms are not parsed from query strings and no implicit operators are used to associate terms. The search system does not make use of spelling error correction or concept-based searching techniques, and search results do not appear to be sorted alphabetically or by relevance.

2.6 Summary

Table 2.1 illustrates the breakdown of food search systems according to the following features: spelling error correction, concept-based searching, search result sorting, and query text parsing. Of the systems reviewed, none possessed a concept-based search and only DietPower 4.4 was capable of detecting and correcting spelling errors in user queries. Additionally, the

DietPower 4.4 SmartSearch made use of limited relevance based sorting, but all other food search systems either sorted results alphabetically or not at all.

	Spelling	Concepts	Sorting	Parsing*
Nutribase EZ Edition 7.10	NO	NO	NONE	YES
DietPower 4.4	YES	SYNONYMS	ALPHA**	YES
Food Processor SQL 10.2.2	NO	NO	NONE	YES
BeNutrifit 1.7	NO	NO	ALPHA	NO
NutriGenie Optimal Nutrition 7.5	NO	NO	NONE	NO
DINE Healthy 7.0	NO	NO	ALPHA	YES

* Indicates that query strings are parsed into individual terms for comparison

** SmartSearch adds additional sorting capabilities

Table 2.1 Food Search System Comparison

CHAPTER 3: Review of Literature Review and Analysis

3.1 Information Retrieval

The focus of this research can be described generally as an experiment evaluating the success of using certain text-based searching techniques in a domain-specific application. As such, it falls under the broader topic of Information Retrieval. The academic field of Information Retrieval may be defined as locating documents or information of an unstructured nature (usually text) that satisfy an information need from within large collections stored on computers (Manning et al. 2008). In this case, documents consist of food item records stored in a relational food database accessed by the DINE Healthy software. Food search users attempt to locate these documents to satisfy an information need. Specifically, users are seeking to add food items to a daily food intake record.

3.1.1 Evaluation of Search Models

Research suggests that most people will not make use of a complicated search interface but will instead give up if the process of searching is too difficult (Search Tools, 2003). Further, a study analyzing the results of a very large web search engine query log (Silverstein et al. 1999) reports that more than 85% of all searches ended after viewing only a single screen. This emphasizes the importance of efficient document searching systems by demonstrating that users often approach searching in a somewhat tentative manner.

With nutrition assessment software in particular, the cost of such behavior could be very high. Failed searches may result in underreporting of food intake or the selection of unacceptable food item substitutions. These outcomes would clearly undermine the value provided by nutrition software programs, and effective search models must be developed to

avoid them. Search model enhancement typically begins by evaluating the existing model to use as a basis for comparison (Hull 1993).

Information retrieval systems are most commonly evaluated in terms of two general measures of effectiveness: precision and recall (Jurafsky and Martin 2000). Precision is a measure of the fraction of returned results that are relevant to the information need and can be thought of as a measure of how much noise is present in the result set. Alternatively, recall is a measure of the fraction of relevant documents in the collection that were returned by the system. The relationship of relevant and retrieved search results to the calculation of precision and recall are illustrated in Figure 3.1.

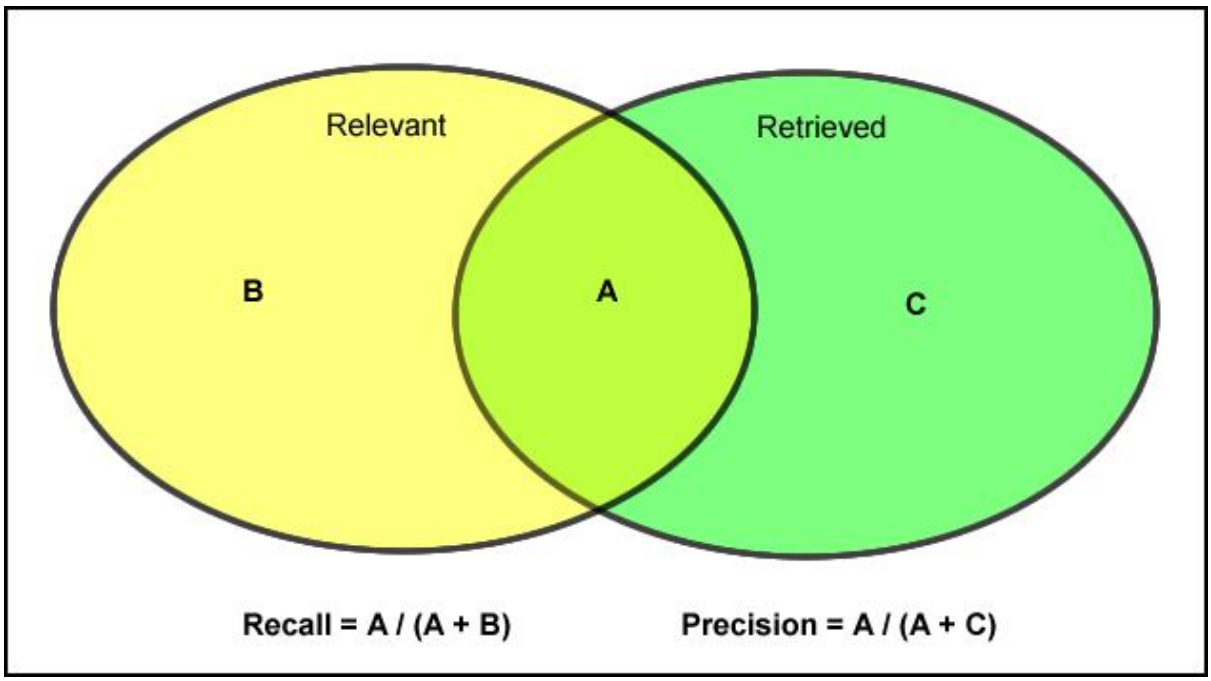

Figure 3.1 Relationship of Recall and Precision

A tradeoff situation exists between these two areas of evaluation, such that a major change in one often results in an inverse reaction with respect to the other (Manning et al. 2008). This

research will focus on improving the recall of the DINE Healthy food search system without causing a significantly negative impact on precision.

3.1.2 TREC and the Cranfield Paradigm

Researchers have been seeking to evaluate Information Retrieval systems through benchmarking since as early as the 1960s when Cyril Cleverdon and colleagues created the first test collection for the Cranfield tests. These experiments lead to the formation of a widely used retrieval system evaluation method known as the Cranfield Paradigm. The Cranfield Paradigm is based on the abstraction of a test collection, a set of documents, a set of information needs, and a set of relevance judgments indicating which documents should be retrieved for each topic (Voorhees, 2005).

Subsequent efforts through the National Institute of Standards and Technology resulted in the formation of the Text REtrieval Conference (TREC) in 1992. TREC was created for the purpose of providing the infrastructure necessary for large-scale evaluation of text retrieval methodologies (TREC, 2008) and assumes the use of the Cranfield Paradigm for experimentation. Despite widespread acceptance in the Information Retrieval community, TREC evaluation methods have still attracted criticism.

One source of concern for critics has been the use of relevance judgments made by human assessors. Objections to this practice stem from the perceived arbitrary nature of human relevance judgments, and some question whether such a metric is valuable. However, research has shown that although relevance scores do vary when different relevance assessors are used, relative scores between groups do not vary significantly (Voorhees, 2005). For example, if system A outperforms system B in the judgment of one assessor, this same conclusion is almost always reached when the two systems are compared by a difference assessor. This finding is

significant in that it supports the effectiveness of human intuition as applied to judging document relevance.

3.1.3 User Tasks and Interactive Searching

Others have cited that the Cranfield Paradigm fundamentally neglects certain aspects of the nature of information seeking, particularly those pertaining to user interaction (Tunkelang, 2008). This includes the goals leading people to engage in information seeking behavior and the tasks associated with these goals. Evidence suggests that these are important because they affect the way people interact with Information Retrieval systems (Belkin, 2008).

To conduct this experiment in a manner consistent as possible with common food search usage scenarios, several techniques are employed to facilitate direct user interaction with the search systems:

- Queries consist of user-formulated search strings, rather than proscribed queries.
- User interaction with the search system is measured through the collection of data such as time spent searching, exact query text entered, and the types and number of sub-queries entered.
- Evaluation metrics are calculated based on user food item selections, rather than only returned search results.

3.2 The Boolean Search Model

The current DINE Healthy food search is based on a Boolean Search model common among consumer data entry applications. Boolean Search systems are capable of posing any query entered in the form of a Boolean expression (Manning et al. 2008), which is a collection of terms combined with the operators AND, OR, and NOT. The DINE Healthy food search parses query strings on space, hyphen, comma, and period characters. Search terms are linked by the AND

operator to construct complex queries. In the DINE Healthy food search, the OR and NOT operators are not implemented.

Implicit operators are often used in Boolean Search systems to simplify the process of entering queries so that all strings need not be interpreted as literals (Jasco 2004). By composing queries that link all search terms with AND operators, the DINE Healthy food search system dictates that food item descriptions must contain all search terms to qualify as a match. This behavior can lead to the exclusion of otherwise acceptable food items from result sets because of misspellings or the use of alternative terms in search queries (such as "diet" rather than "low calorie") that do not explicitly match food descriptions.

Another important feature of the Boolean Search model is that results are not ranked by relevance but simply either match or fail to match the query (Manning et al. 2008). In this type of search model, the result set can either be returned entirely unsorted, or it can be ordered by some other means. For example, DINE Healthy food search results are ordered alphabetically by food item description. The noun most generally representing each food item occurs first in its description in the DINE Healthy food database, thus alphabetical ordering was chosen as the most reasonable method for sorting results in lieu of relevance.

3.3 Spelling Error Correction

The observed rate of spelling errors affecting the retrieval of documents from search engines is quite high, with one study reporting a rate of over 26% (Wilbur et al. 2006). Because these errors occur so frequently in search engine queries, it is reasonable to suspect that a significant number of spelling errors are also present in DINE Healthy food search queries. (In fact, search log data collected from subjects over the course of this research does show that a significant number of queries contain spelling errors: over 34%). This presents a major obstacle for DINE

Healthy users because the current food search cannot effectively produce results for misspelled query terms. No results are returned if any term in the user query contains a misspelling or a word not in the database.

3.3.1 Subtask Classification

Much research exists documenting the various causes of and resolutions for spelling errors. The problem of detecting and correcting these errors is commonly divided into three subtasks (Kukich 1992) of increasing difficulty:

1) non-word error detection,

2) isolated-word error correction, and

3) context-dependent error correction.

Non-word error detection can be accomplished relatively easily by comparing each query string term with a known dictionary, such as the set of all words in the document collection. Terms can then be identified as misspellings if they are not located in the dictionary. Very low frequency words often have a high probability of being misspelled forms of other words (Wilbur et al. 2006), and it may be appropriate to suggest corrections for terms falling below a specified frequency threshold by offering to substitute other terms of higher frequency.

Isolated word error correction is the process of correcting spelling errors that result in non-words and is accomplished by looking only at the word in isolation. Single word queries are devoid of context, and thus non-words of this sort are subject to isolated-word error correction. Context-dependent error correction is useful in the detection and correction of spelling errors resulting in actual English words, known as real-word errors (Jurafsky and Martin 2000). Unfortunately, most search queries consist of a maximum of two or three words (Silverstein et al. 1999), and this can limit the effectiveness of a context-dependent approach.

3.3.2 The Noisy Channel Approach

The noisy channel (or Bayesian inference) approach to spelling correction consists of proposing and scoring candidate corrections according to edit distance and probability. To limit the set of terms for which edit distances must be calculated, a k-gram index can be used to locate candidate corrections that have many k-grams (or k-length substrings) in common with the misspelled term.

Levenshtein distance is the number of edit operations (insertions, deletions, or character substitutions) required to transform a term into its lexical form (Manning et al. 2008) and is commonly used to represent edit distance. As shown in Table 3.2, the Levenshtein distance of the string “Cantelope” from “Cantaloupe” is 2.

		C	a	n	t	e	l	o	p	e
	0	1	2	3	4	5	6	7	8	9
C	1	0	1	2	3	4	5	6	7	8
a	2	1	0	1	2	3	4	5	6	7
n	3	2	1	0	1	2	3	4	5	6
t	4	3	2	1	0	1	2	3	4	5
a	5	4	3	2	1	1	2	3	4	5
l	6	5	4	3	2	1	1	2	3	4
o	7	6	5	4	3	3	2	1	2	3
u	8	7	6	5	4	4	3	2	2	3
p	9	8	7	6	5	5	4	3	2	3
e	10	9	8	7	6	5	5	4	3	2

Table 3.1 Levenshtein Distance Example

3.4 Concept-Based Searching

There are advantages to conducting searches in terms of concepts rather than specific words. A single subject can often be denoted by a number of different words (synonymy), which badly undermines the effectiveness of the Boolean Search model. Under these circumstances, word-based queries fail to retrieve relevant documents lacking the specific terminology required to satisfy the query. Most concept-based searching techniques attempt to overcome this limitation by introducing domain knowledge directly into the search model in the form of a domain-specific ontology (Clark et al. 2000). This approach requires the construction of a customized conceptual vocabulary to link concepts by relationship such as “broader term”, “narrower term”, and “related to”.

3.4.1 Ontology Creation

An ontology is a formal representation of the set of concepts and the relationships between those concepts within some domain. They are commonly constructed to share understanding of the structure of information, to capture and make domain knowledge available for reuse and to make domain assumptions explicit (Noy & McGuinness, 2001). Unfortunately, the high up-front costs associated with ontology development have historically hindered research attempts in this area (Clark et al. 2000). As the DINE Healthy software currently contains a database of over 10,000 individual food items, the task of creating an adequate conceptual vocabulary for the entire database would prove extremely challenging.

One approach to overcoming the development time obstacle would be to make use of a pre-constructed ontology of food terms, such as the Food Ontology Project maintained by the Genomic Standards Consortium (GSC 2008). This would likely be effective if entries are highly correlated with terms used in DINE Healthy food item descriptions. However, a significant risk

of reducing precision exists if the ontology and food database vocabulary are not highly related (Ide et al. 2007). An alternative approach is to create a less comprehensive ontology detailing concept relationships for a subset of high-frequency terms.

3.4.2 Ranking Results by Relevance

The ability to overcome the semantic limitations of the Boolean Search model is not the only benefit achieved by concept-based searching. Another advantage of this technique is that it provides additional information to the search system required to successfully rank search results by relevance (Ide et al. 2007). This is in contrast to the current DINE Healthy food search that ranks search results alphabetically.

CHAPTER 4: Implementation of Search Enhancements

4.1 Spelling Error Correction

The first experimental food search version developed for use with the DINE Healthy software adds automatic spelling error detection and correction of text-based queries. As with most spelling error detection systems, misspellings are identified by consulting a dictionary of known terms and frequencies of occurrence. A publicly available dictionary could potentially be used for this purpose, but because of the heavily domain-specific nature of the software, a dictionary of terms and frequencies is created from food item descriptions in the DINE Healthy food database. Once identified, misspellings are automatically replaced with the most likely candidate term in the dictionary.

4.1.1 Dictionary of Terms and Frequencies

The first step in this process is to create a dictionary of known terms and their frequency of occurrence. This is accomplished by iterating through the 10,156 food item records in the DINE Healthy food database and parsing each of the corresponding description strings into individual terms. The method used to parse food item descriptions is extremely important because user entered queries must be parsed in exactly the same way. The set of delimiting characters are shown in Table 4.1. For each term in the dictionary, an occurrence count is maintained and incremented each time it is found in a food item description.

space	comma	backslash	open parenthesis	close parenthesis	hyphen	ampersand
	,	/	(	)	-	&

Table 4.1 Term Delimiters

A dictionary created from the set of all food item descriptions in the DINE Healthy food database contains 3,700 unique terms with a mean occurrence count of 19. Table 4.2 shows a sample of dictionary terms and their corresponding occurrence counts.

Term	Count	Term	Count	Term	Count
abalone	2	baby	297	california	14
abc's	4	bacardi	10	calistoga	1
abiyuch	1	bachman	3	calorie	39
accents	4	back	12	calzone	2
access	1	bacon	63	camembert	1
acerola	2	bacos	1	cameo	1
acesulfame	5	bag	13	campbell's	50
acid	7	bagel	52	can	14
acorn	2	bagelette	3	canada	501
acorns	1	bake	21	canadian	8
act	12	baked	124	candied	4
acting	2	baker's	2	candies	5
active	1	bakery	19	candy	149
activia	3	bakes	2	canned	409
added	87	baking	39	cannelloni	2
adherent	1	baklava	1	cannoli	1
adobo	1	balance	17	canola	12

Table 4.2 Sample Dictionary Terms

4.1.2 Spelling Error Correction Algorithm

The spelling error correction algorithm used in this research is a straightforward interpretation of the Noisy Channel (or Bayesian Inference) model and is used to perform isolated-word error correction on non-words. Context-dependant error correction is not implemented because most queries in this domain are believed to lack sufficient context for the useful application of such techniques. (An examination of all text-based queries collected during the course of the experiment reveals that 71% consist of a single term, 25% two terms, and only 4% have more than 2 terms).

The first step in the process of correcting a spelling error is to identify that a non-word situation exists. Specifically, it is necessary to identify a term as a non-word if it is not found within the domain terminology. This is accomplished by consulting the dictionary of terms found in the corpus (i.e. the DINE Healthy food database). If a term is not found in the dictionary, it is identified as a non-word and the system attempts to locate the most probable correction.

The task of locating candidate corrections begins with calculating the Levenshtein distance between the non-word and each dictionary term. Levenshtein distance is calculated by summing all the text operations (i.e. insertions, deletions, and substitutions) minimally required to transform one string into another. For simplicity, a weighting factor of 1 is assigned as the cost of each of these operations. Because of the relatively small dictionary (3,700 unique terms), this can be accomplished very quickly.

The set of dictionary terms having the smallest distance from the non-word are selected as candidate corrections. For each candidate correction, the probability is calculated by dividing the total number of occurrences of that term by the total number of occurrences of all terms in the corpus. The candidate with greatest probability is selected as the non-word error correction. Table 4.3 shows the candidate corrections generated for the non-word "ric". In this example, the edit distance for each candidate is 1, and "rice" has the greatest probability of occurrence.

Candidate	Distance	Probability
rice	1	0.004048
rib	1	0.001919
rich	1	0.000378
rio	1	0.000028
rc	1	0.000014

Table 4.3 Candidate Corrections for Non-Word ("ric")

4.1.3 Enhanced Food Search Algorithm

When the user enters a text-based query, it is parsed into individual terms. Terms that do not provide any discriminating value (i.e., *the, a, of, an,* etc.) are removed from the query automatically. The remaining terms are checked to determine if they exist within the corpus. If a term is classified as a non-word, the spelling error correction algorithm automatically replaces it with the most likely candidate correction.

The query text displayed in the "Search" textbox is updated to reflect any spelling error corrections. This is done to reduce the chance that incorrect term replacements go unnoticed by the user and to enable further adjustments to be made to the query if necessary. Results are then generated by matching the list of terms with food item descriptions. Ranking is done alphabetically (rather than by relevance as with the concept-based search discussed in Section 4.2). Figure 4.1 shows the result of entering the misspelled query "brocoli speares" which is automatically corrected to "broccoli spears". Note that the "Search" textbox is updated.

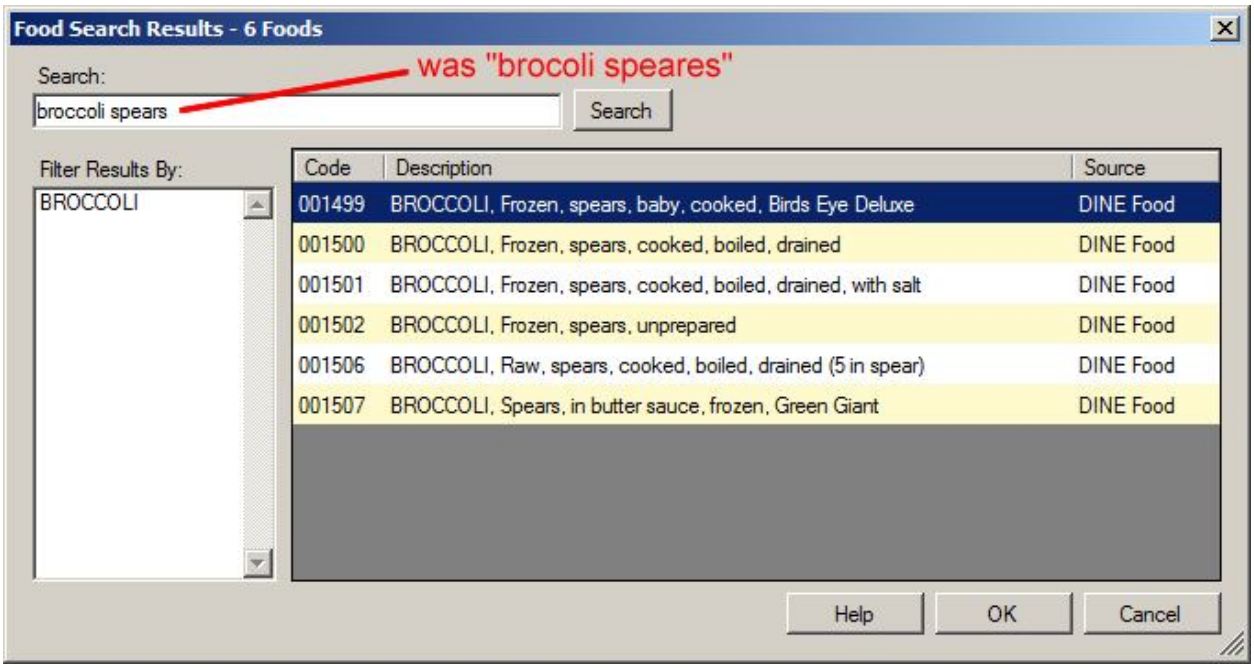

Figure 4.1 Food Search Results with Spelling Error Correction

4.2 Concept-Based Searching

A second experimental food search version is created by adding concept-based searching to the enhanced food search with spelling error correction. The development of a concept-based search requires that domain knowledge be captured and made available to the search in a usable fashion. In this experiment, such knowledge is leveraged for the purpose of expanding user queries and sorting food search results by relevance.

However, obtaining domain knowledge is often a difficult task. Even specifically defining the domain boundaries can be difficult. Existing human food ontologies are available, but these tend to focus on nutritional information rather than concepts well suited for use with food searching (GSC 2008). Pertinent knowledge can be extracted from the DINE Healthy food database, but the task of creating a complete ontology for all 10,156 food items is quite daunting. For these reasons, the scope of the ontology is limited by focusing on the 20 food items used in the experiment.

4.2.1 Gathering Domain Knowledge

To elicit usable domain knowledge from these 20 food items, a concept map is constructed for each. Concept maps consist of two basic elements; concepts and the relationships linking those concepts. In this domain, concepts represent ideas associated with food items such as their description, preparation, manufacturer, etc. Relationships provide insight into the ways in which concepts interact with one another.

To construct each map, a food item is used as the central concept to which others are attached by various relationships. Each relationship has a source and target concept and can be directional or bidirectional in nature. With bidirectional relationships, the source and target concepts are effectively reversible. Another way of stating this is that an implied, identical

relationship exists in which the source and target concepts are reversed. Table 4.4 shows the seven individual relationships modeled in this experiment. It is important to note that numerous relationships may potentially exist within this domain, but these seven are collected because they provide insight and are relatively easy to identify.

Relationship	**Example**	**Bidirectional***
stem_of	candy - stem_of - candies	No
synonym_of	catsup - synonym_of - ketchup ketchup - synonym_of - catsup	Yes
very_like	chopped - very_like - cubed cubed - very_like - chopped	Yes
is_a	apple - is_a - fruit	No
main_ingredient	egg - main_ingredient - omelet	No
prepared	broccoli - prepared - chopped	No
sells	Wendy's - sells - watermelon	No
* An identical relationship is implied from the target concept back to the source		

Table 4.4 Set of Relationships

Developing a concept map begins with defining the central concept. This is typically a general term describing the food item to be mapped (e.g. "tomato"). Food items in the DINE Healthy food database that represent a specific instance of the central concept are then analyzed. Concepts are selected from these food item descriptions if they are associated with the central concept by one of the seven relationships.

Additionally, food item descriptions for similar or related concepts are examined to extract any potentially useful information. For example, "catsup" is added to the concept map for "tomato" because tomatoes are the main ingredient in catsup. This is something of a best effort method, in that no process exists for exhaustively capturing all relationships. However, care is

taken to represent each of the food items adequately within the scope of the domain. Figure 4.2 shows a concept map constructed for “tomato”.

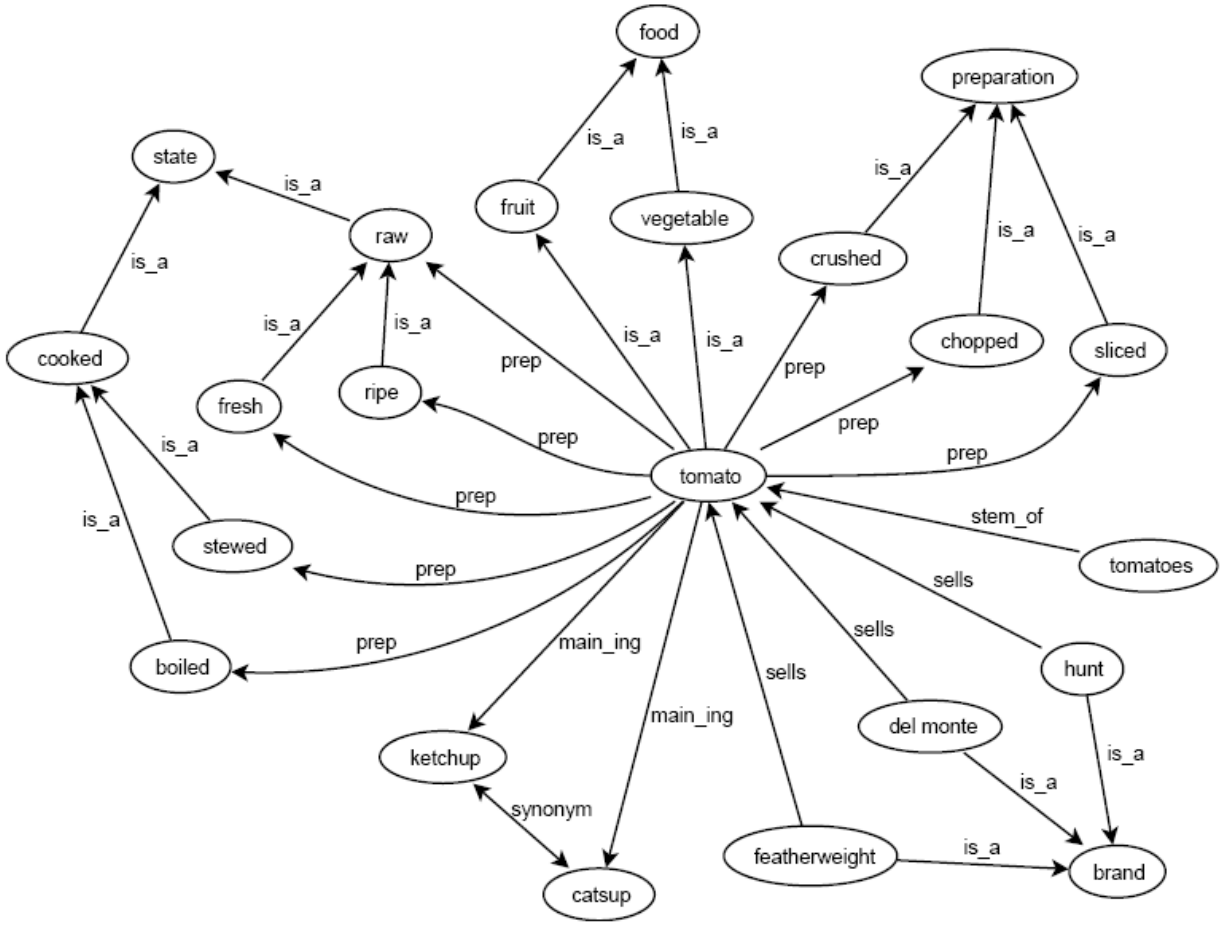

Figure 4.2 Concept Map for “tomato”

4.2.2 Ontology and Query Expansion

Constructing an ontology from a set of concept maps requires combining the full set of maps into a single, structured repository of knowledge. When the individual concepts are aggregated, this immediately exposes the existence of several high-level concepts that together categorize most others (i.e., food, state, brand, and preparation). This information is useful because it explicitly identifies the more general concepts that users might wish to express in their queries.

For example, a text-based query of “raw tomatoes” is identifiable as an effort to locate a food (tomato) in a particular state (raw) because of these general classifications.

A primary function of the ontology is to facilitate query expansion: the process by which text-based queries are generalized from words into concepts by expanding their individual terms into weighted lists of terms. Query expansion depends on the ability to measure the relevance of concepts to one another within the ontology. For this purpose, each relationship is assigned a numeric weight representing the strength of the association, or the a-priori probability of relevance. Research suggests that relevance weights need not approximate actual probabilities so long as the relative ranking between them is preserved (Ide et al, 2007). For example, stems of a given term (i.e. chop, chopped, chops, chopping) are assigned the highest weight (0.95) because they are nearly identical conceptually.

Each relationship also defines the direction(s) in which the expansion algorithm can proceed. This is because some relationships overgeneralize or are simply inappropriate paths for query expansion. This scenario is apparent in the “is_a” relationship, which is expanded from target to source only (e.g., a search on vegetables should include tomatoes, but a search on tomatoes should not include all vegetables). Table 4.5 lists the weights assigned to each of the seven relationships and the directions in which query expansion is possible.

Relationship	Weight	Expansion
stem_of	0.95	bidirectional
synonym_of	0.75	bidirectional
very_like	0.60	bidirectional
is_a	0.50	toward source
main_ingredient	0.30	toward target
prepared	0.15	none
sells	0.15	none

Table 4.5 Relationship Weighting Scheme

4.2.3 The Query Expansion Algorithm

The enhanced food search process begins when a text-based query is entered and parsed into a set of individual terms. Each term is checked against the dictionary (which includes all concepts in the ontology), and non-words are replaced with the most probable corrections. The remaining terms are analyzed for adjacent pairs that together represent individual two-word concepts in the ontology (e.g. “soda pop”, “soft drink”, “egg substitute”, etc). Because they are effectively just components of one concept, these pairs are combined into a single term. This process is illustrated for the query “diet grape soft drink” in Figure 4.3.

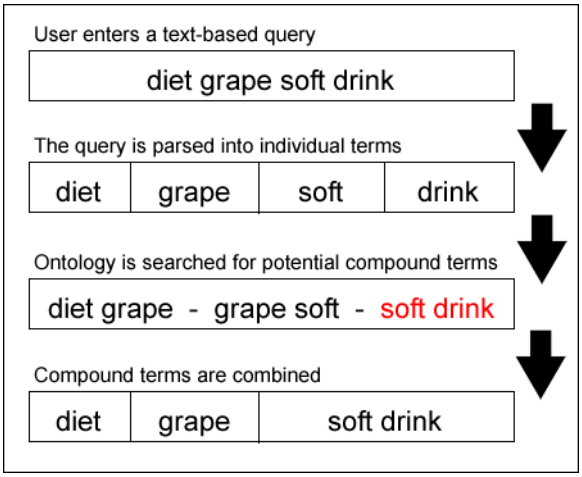

Figure 4.3 Term Compounding Example

Prior to expansion, each term is converted into a weighted list containing only the original term with a weight of 1.0. The search proceeds by iterating through each list in the set and calling the query expansion algorithm. This recursively expands each list by adding related concepts until no more are found. Concepts are only added to the list under the following circumstances:

- The concept has not already been added to the list by way of another relationship.

- The connecting relationship permits expansion in the direction of the associated concept as indicated in Table 4.5.
- The calculated relevance weight exceeds the minimum threshold (0.1)

Relevance weights are calculated by multiplying the relevance weight of the listed concept by the relationship weight connecting the associated concept. This provides an easily quantifiable measure of distance between related concepts while creating an automatic self-limiting boundary for the expansion algorithm. The effect is that the relevance metric for each added concept decreases in proportion to the relevance lost through all relationships in the path connecting it to the original concept.

Figure 4.4 illustrates how the relevance weights are calculated for a given set beginning with the original concept “potatoes”.

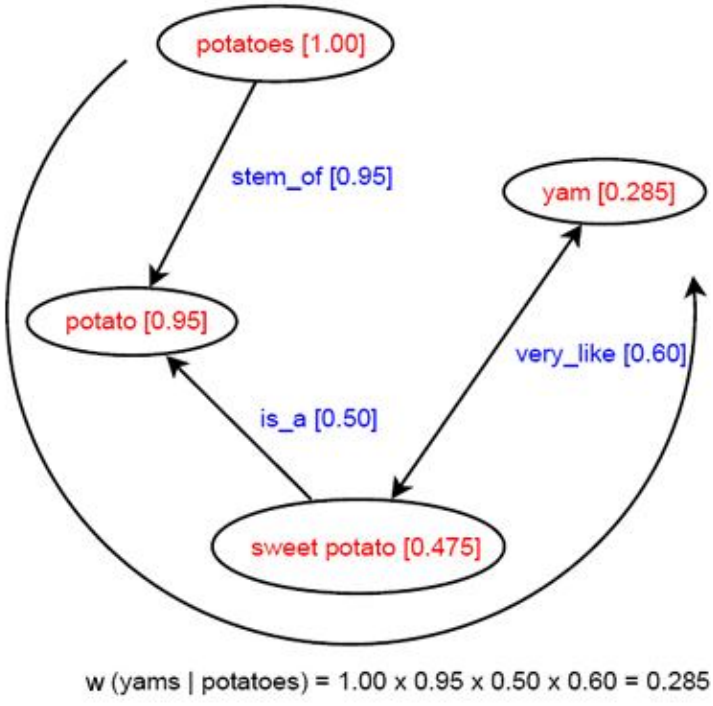

Figure 4.4 Example Calculating Relevance Weights

4.2.4 Sorting Results by Relevance

Unlike other search versions which rank results alphabetically, the concept-based food search orders results by relevance. This is accomplished using a dynamic programming algorithm which is probably easiest to visualize from the top down.

- At the outer layer, a relevance score (ranging from zero to one) is calculated for each food item in the DINE Healthy food database. This is obtained by calculating the mean score for the set of all weighted term lists generated by the query expansion algorithm. Food items with a score greater than zero for each weighted term list are sorted and returned as results.
- The score for a weighted term list (ranging from zero to one) is obtained by calculating the maximum score of all terms in the list.
- This leads us to the process of calculating the score for an individual term, which is the basic unit of the relevance scoring algorithm.

The term scoring procedure begins by attempting to locate the term in the food item description and returning a score of zero if it is not found. If the term is located, the position of the term (p) is calculated as the index of the leftmost character of the term as it appears in the description. Food item descriptions may be up to 256 characters in length, thus the range for this value is from zero to 255.

DINE Healthy food item descriptions are structured such that concepts most generally representing the food item are located at or very near the beginning of the description. The proximity weighting function shown in Figure 4.5 is designed to take advantage of that structure by assigning a high proximity weight to such terms. This function is used to calculate the relevance score for the term from position (p).

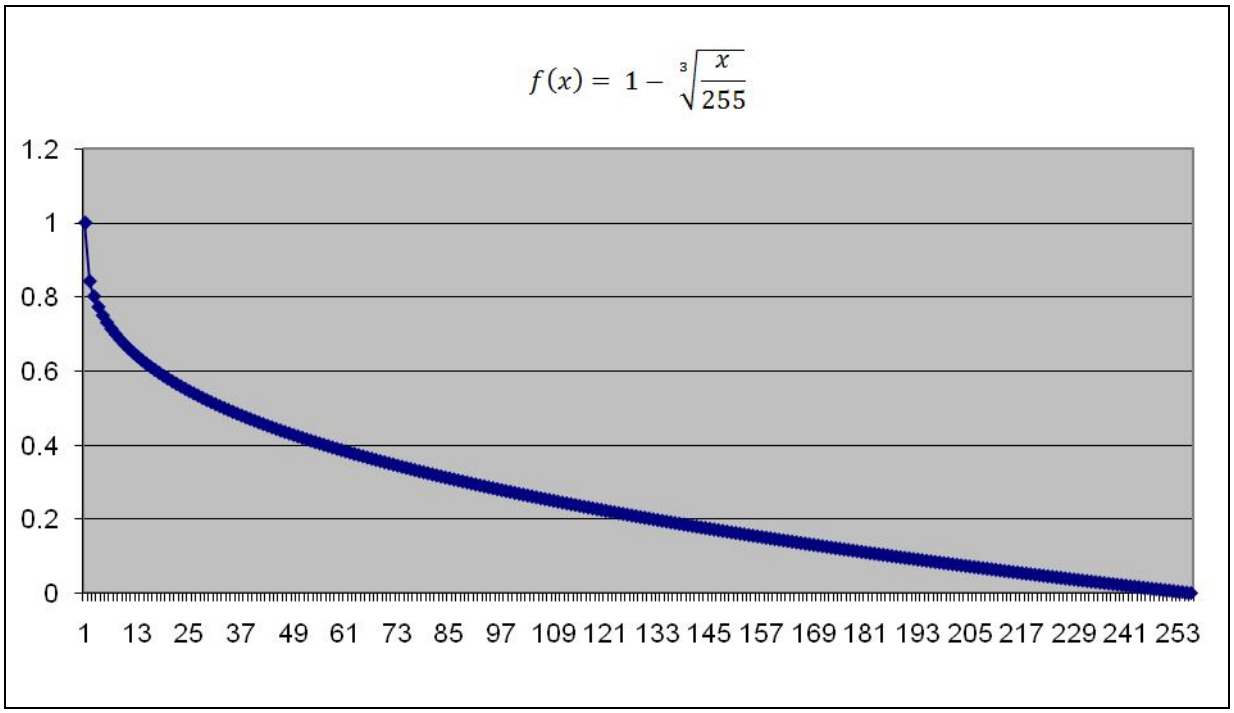

Figure 4.5 Proximity Weighting Function

It bears mentioning that no assumption is made that the proximity weighting function used in this research is the ideal function for this purpose, or even for this domain. The choice of the specific function is based on the fact that it conveniently generates a curve assigning disproportionately greater value to terms occurring at or very near the beginning of the food item description. Other functions (including linear, square root, and 4th root forms of the same function) have been tested but do not show equally promising results.

CHAPTER 5: The Experiment

5.1 Overview

The primary research objective, to measure the effects of enhancing the DINE Healthy food search with advanced searching techniques, is accomplished by conducting an experiment in which usage data is collected from three groups of human subjects and analyzed to evaluate the performance of each search version. The key tasks required to conduct this experiment are performed as follows:

1) Develop a client server data logging system to collect detailed information about DINE Healthy food searches.
2) Create a packet of full page color images representing common food items.
3) Collect data from a control group instructed to locate each pictured food item using the current DINE Healthy food search.
4) Develop two additional food search versions for the DINE Healthy software; one extending the search with spelling error correction, and the other extending the search with both spelling error correction and concept-based searching.
5) Collect data from two experimental groups, each using the same food item image packets and one of the extended search versions.
6) Parse and analyze search log entries.

5.2 Data Collection

When performing a DINE Healthy food search, users may choose to initiate various combinations of actions which include entering text-based queries, filtering by category, and selecting food items. All of these constitute user interaction with the search system and should be recorded. For this reason, a flexible representation of food searches is needed to fully capture

user searching activity. The first step toward creating this representation is to define the scope and structure of a food search event.

Start and end times mark the beginning and ending boundaries of a food search. Start time is captured when the Food Search Results window is initialized in response to a user search request (i.e. pressing the enter key in a search field or clicking the "Search" button). End time is captured when the window is closed by the user. This occurs after a food item has been selected or the user chooses to "Cancel" the search without making a selection. Everything that occurs between the start and end times, including text queries, category filtering and food item selection, composes a single food search event.

Text-based queries and category filtering operations are treated as individual query events occurring within the search, and several metrics are recorded for each:

- Sequence number indicating the order in which each query occurs
- Type of query, as in "TEXT" for text-based queries and "FILTER" for category filtering operations
- Text entered by the user for a text query, or the category filter selected
- Number of results generated by the search system in response to query
- Timestamp indicating when the query was initiated

The user selected food item is the final component of the search event, and a list of food item selections is recorded for each. This list may contain zero, one, or multiple selections. An integer uniquely identifying the food item in the DINE Healthy food database and a text description of the food item are recorded.

As each food search event is completed, the DINE Healthy software automatically encodes the collected search log information in an XML document and transmits it to a log server via

TCP/IP connection. Figure 5.1 shows a sample XML log entry for a single food search. In this example the subject is attempting to locate “TOMATOES, Raw, (2-3/5 in diameter)” by searching on the query string “tomato”, followed by “raw”, then filtering on “TOMATOES”. The total search time 21 seconds.

```
<?xml version="1.0" encoding="utf-16"?>
<search>
  <start>3/24/2009 12:43:03 PM</start>
  <end>3/24/2009 12:43:24 PM</end>
  <queries>
    <query>
      <sequence>0</sequence>
      <type>QUERY</type>
      <text>tomato</text>
      <results>219</results>
      <timestamp>3/24/2009 12:43:03 PM</timestamp>
    </query>
    <query>
      <sequence>1</sequence>
      <type>QUERY</type>
      <text>raw</text>
      <results>455</results>
      <timestamp>3/24/2009 12:43:10 PM</timestamp>
    </query>
    <query>
      <sequence>2</sequence>
      <type>FILTER</type>
      <text>TOMATOES</text>
      <results>12</results>
      <timestamp>3/24/2009 12:43:17 PM</timestamp>
    </query>
  </queries>
  <selections>
    <selection>
      <foodid>9266</foodid>
      <description>TOMATOES, Raw (2-3/5 in diameter)</description>
    </selection>
  </selections>
</search>
```

Figure 5.1 Sample XML Search Log Entry

5.3 Food Item Images

The food search model evaluation method developed for this experiment measures performance based partly on the success of subjects using the search system to accomplish specific goals. This method has the benefit of incorporating the user's ability to cope with the search system into the evaluation process. However, in order to accomplish this, a uniform set of search goals must be established for all subjects. As research suggests that the goals leading people to engage in information seeking behavior, and the tasks associated with those goals affect the way they interact with the system (Belkin, 2008), care must be taken to develop a set of tasks that are both realistic and within proportion to those of typical users.

In addition to these constraints, factors specifically affecting the function of the evaluated searching techniques must be considered. For example, providing subjects with instructions to locate food items based on a list of text descriptions would be unacceptable. Such a practice would severely limit the potential impact of spelling error correction by providing food item description spellings automatically.

Further, the use of recorded audio instructions would provide too much coaching and inhibit subjects from forming their own queries. For example, instructing subjects to locate "grape soft drink" would likely impact the behavior of some subjects that may otherwise enter the query "grape soda". In order to examine the effectiveness of concept-based searching, user queries must be free from such influences.

For these reasons, a set of 20 full page color food item images are used as search goals. Care is taken with each image to display food items in serving sizes and presentations that are familiar to most subjects. A wide variety of foods are represented (e.g. beverages, fast foods, vegetables, fruits, breads, and pastas) to approximate the types of entries made by typical users

of the DINE Healthy software. Figure 5.2 shows a sample food item image (STRAWBERRIES, Raw) used in the experiment.

Figure 5.2 Sample Food Item Image

5.4 Human Subjects

Data examined in this research are collected from individual subjects using the DINE Healthy software in a computer lab setting. As with any research conducted using human

subjects, all reasonable effort must be made to ensure they are protected from harm. No personally identifiable data are collected from any subject and the risks associated with participating in the experiment are believed to be the same as those typically encountered in everyday life. Informed written consent is given by all subjects prior to their participation.

5.4.1 Sample Population

Human subjects are drawn from a sample population of UNCW students enrolled in HEA 207 and HEA 359 during the Spring Semester of 2009. The sample consists of 34 students randomly selecting into three groups. A control group of 11 subjects is selected and tasked with locating food items using the current DINE Healthy food search. An experimental group of 10 students is tasked with locating food items using the enhanced search with spelling error correction, and a second control group of 13 students is tasked with locating food items using the enhanced search with both spelling error correction and concept-based searching. The sample population consists of both males (24%) and females (76%) between 18 and 30 years of age.

5.4.2 Lab Setting and Preparation

Food item search tasks and data collection are carried out in the Trask Computer Lab on the UNCW campus. Subjects are invited to participate by way of an email announcing the location, date, and time that information is to be collected. When a subject arrives at the testing location, he or she receives a consent form explaining the basic elements of user participation in the experiment. Subjects are aware that log data containing search times are collected, but the specific use of this data is not mentioned to prevent any impact on typical food searching behavior.

Once consent is obtained, the subject is seated at a computer terminal with the DINE Healthy software running and a daily Food Record window open. He or she receives a packet of

20 food item images and is instructed to locate each one and enter it into the Food Record. Prior to starting the session, the administrator demonstrates how the food search interface is used to locate a single food item (i.e. “Apples, Raw w/out skin (2-3/4 in diameter)”). The subject then attempts to locate and enter each food item from the packet. When the subject has entered all foods that he or she is able to locate, the administrator closes the Food Record window and the session is complete.

CHAPTER 6: Experimental Results

6.1 H_A 1: Reduction of Mean Search Time

"Subjects using the enhanced food search models will take less time to complete the assigned food searches than subjects using the current DINE Healthy food search."

To calculate mean search time, the following factors must be considered:

- Some attempts to locate a single food item span multiple searches.
- Not all searches end in a food item selection.

In light of these facts, a simple evaluation of mean seconds per food search could be misleading. This is because several individual searches may actually be components of a single food item seeking event and should not be compared separately with respect to time. Further, limiting the scope of analysis to only successful searches would neglect time spent that did not directly result in a selection. For these reasons, evaluation is based on the mean number of seconds spent searching per food item selection.

6.1.1 The Spelling Only Group

Table 6.1 shows the results of a Two-Sample t-Test comparing mean search time per selection for the Control and Spelling Only groups. N is equal to the number of food item selections made by all members of the indicated group.

Group	**N**	**Mean (sec.)**	**STDV**
Control	218	31.36	35.1
Spelling Only	200	20.86	22.2
Test	**DF**	**T**	**P**
Two-Sample t-Test	371	3.69	< 0.0002

Table 6.1 t-Test of Mean Search Time for Spelling Only Group

Data collected from the Control group consists of 218 observations with a mean search time of 31.36 seconds. The Spelling Only group data consists of 200 observations with a mean search time of 20.86 seconds. The difference between means for the Spelling Only group and the Control group is statistically significant ($t = 3.69$, $df = 371$, $p < 0.0002$), which demonstrates that subjects using the enhanced food search with spelling error correction spent less time searching for food items than subjects using the current DINE Healthy food search.

6.1.2 The Spelling + Concepts Group

Data collected from the Spelling + Concepts group consists of 260 observations with a mean search time of 14.93 seconds. The difference between means for the Spelling + Concepts and Control groups is statistically significant ($t = 6.39$, $df = 290$, $p < 0.0001$). This demonstrates that subjects using the enhanced food search with both spelling error correction and concept-based searching spent less time searching for food items than subjects using the current DINE Healthy food search.

Table 6.2 shows the results of comparing mean search time per selection for the Control and Spelling + Concepts groups.

Group	**N**	**Mean (sec.)**	**STDV**
Control	218	31.36	35.1
Spelling + Concepts	260	14.93	15.8
Test	**DF**	**T**	**P**
Two-Sample t-Test	290	6.39	< 0.0001

Table 6.2 t-Test of Mean Search Time for Spelling + Concepts Group

A Two-Sample t-Test reveals that the mean of 14.93 seconds per food item selection observed in the Spelling + Concepts group is significantly different from the mean of 20.86 seconds observed in the Spelling Only group ($t = 3.2$, $df = 344$, $p = 0.0007$). This demonstrates that value is achieved by the addition of concept-based searching to the enhanced food search with spelling error correction.

Table 6.3 summarizes the comparison of means between the Spelling Only and Spelling + Concepts groups.

Group	**N**	**Mean (sec.)**	**STDV**
Spelling Only	200	20.86	22.2
Spelling + Concepts	260	14.93	15.8
Test	**DF**	**T**	**P**
Two-Sample t-Test	344	3.20	0.0007

Table 6.3 t-Test of Mean Search Time for Spelling Only and Spelling + Concepts

6.1.3 Conclusions

Analysis of the experimental results demonstrates that for groups using both enhanced food search versions, mean search time per food item selection is improved by a statistically significant amount. These results represent an improvement in mean search time over control of 10.5 seconds, or 33.5%, using the enhanced food search with spelling error correction. An improvement in mean search time of 16.43 seconds, or 52.5%, is realized using the enhanced food search with both spelling error correction and concept-based searching. Additionally, a significant improvement in mean search time is also shown to result from the addition of concept-based searching to the enhanced food search with spelling error correction.

Figure 6.1 illustrates the improvement with respect to mean search time achieved with both enhanced food search versions.

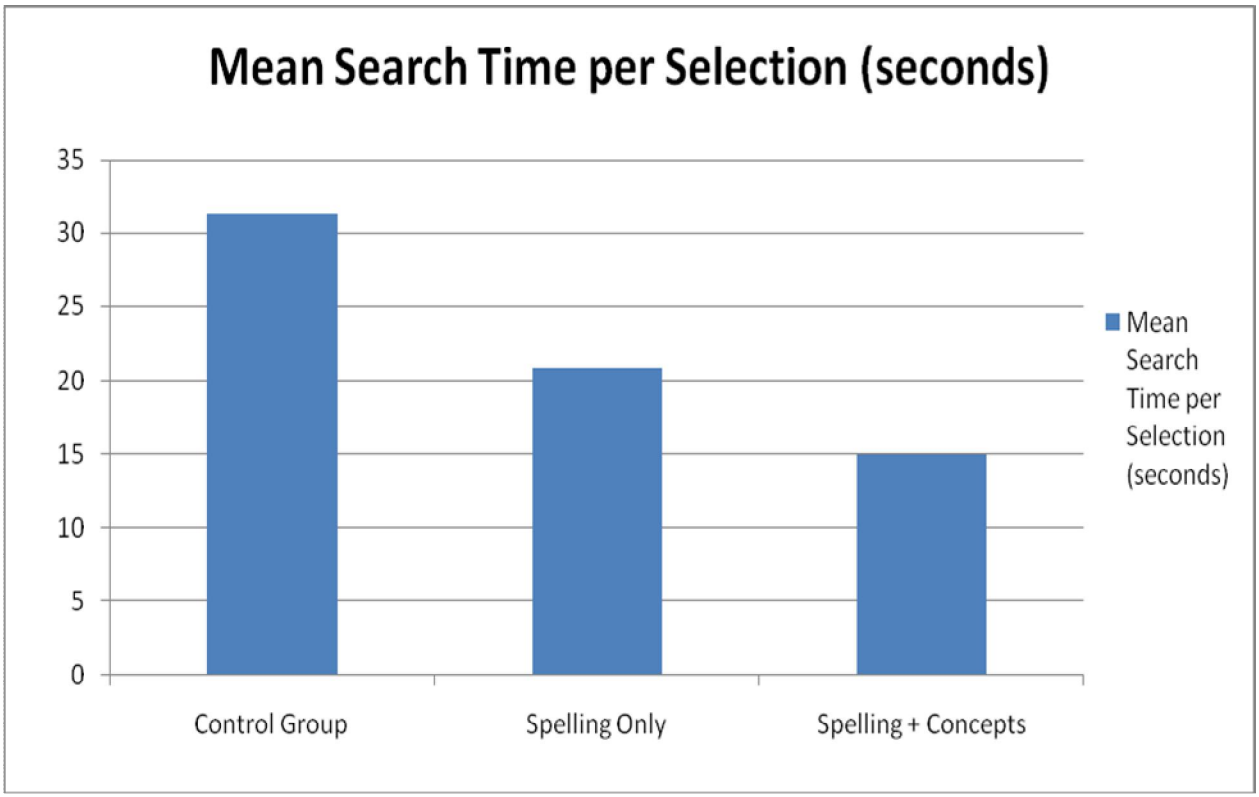

Figure 6.1 Mean Search Time per Selection (seconds)

6.2 H_A 2: Reduction of Failed Searches

"Subjects using the enhanced search models will initiate fewer searches that fail to return an acceptable selection than subjects using the current DINE Healthy food search."

For the purpose of comparing the success or failure of searches, food item selections are classified as follows:

- Correct – the food item selected is among a group pre-specified as a correct selections for the given food item image.

- Acceptable – the food item selected is not among the group pre-specified as correct selections. However, upon inspection it could not be positively eliminated as the pictured food item.
- Incorrect – the food item selected was obviously not the pictured food item.
- Gave Up – no food item was selected.

In this experiment, the group of successful food searches is composed of searches with correct or acceptable selections. Failed food searches are made up of searches with incorrect selections and searches in which no selection was made. This equates to 73 failed searches for the Control group, 13 for the Spelling Only group, and 20 for Spelling + Concepts group.

Table 6.4 shows the breakdown of successful and failed searches into their constituent parts.

	Successful Searches			Failed Searches			Total
Group	Correct	Acceptable	Total	Incorrect	Gave Up	Total	Searches
Control	181	25	206	12	61	73	279
Spelling Only	172	27	199	1	12	13	212
Spelling + Concepts	234	20	254	6	14	20	274

Table 6.4 Breakdown of Food Search Success

6.2.1 The Spelling Only Group

Data collected from the control group consists of 279 observations with a mean failure rate of 26.16%. The Spelling Only group data consists of 212 observations with a mean failure rate of 6.13%. The difference between means for the Spelling Only and Control groups is statistically significant ($t = 6.44$, $df = 448$, $p < 0.0001$), which demonstrates that subjects using the enhanced food search with spelling error correction initiate fewer searches that fail to return an acceptable result than subjects using the current DINE Healthy food search.

Table 6.5 shows the results of a Two-Sample t-Test comparing the percentage of search failures for the Control and Spelling Only groups. N is equal to the total number of food searches conducted by subjects in each group.

Group	N	Mean (%)	STDV
Control	279	26.16	44.0
Spelling Only	212	6.13	24.0
Test	**DF**	**T**	**P**
Two-Sample t-Test	448	6.44	< 0.0001

Table 6.5 t-Test of Search Failure Rate for Spelling Only Group

6.2.2 The Spelling + Concepts Group

Table 6.6 shows the results of a Two-Sample t-Test comparing the percentage of search failures for the Control and Spelling + Concepts groups.

Group	N	Mean (%)	STDV
Control	279	26.16	44.0
Spelling + Concepts	274	7.30	26.1
Test	**DF**	**T**	**P**
Two-Sample t-Test	453	6.14	< 0.0001

Table 6.6 t-Test of Search Failure Rate for Spelling + Concepts Group

Data collected from the Spelling + Concepts group consists of 274 observations with a mean failure rate of 7.3%. The difference between mean search failure rate for the Control and Spelling + Concepts groups is statistically significant ($t = 6.14$, $df = 453$, $p < 0.0001$). This

demonstrates that subjects using the enhanced food search with both spelling error correction and concept-based searching initiate fewer searches that fail to return an acceptable result than subjects using the current DINE Healthy food search.

Table 6.7 shows the results of a Two-Sample t-Test comparing the percentage of search failures for the Spelling Only and Spelling + Concepts groups

Group	**N**	**Mean (%)**	**STDV**
Spelling Only	212	6.13	24.0
Spelling + Concepts	274	7.30	26.1
Test	**DF**	**T**	**P**
Two-Sample t-Test	469	-0.51	0.3046

Table 6.7 t-Test of Search Failure Rate for Spelling Only and Spelling + Concepts

This test reveals that the mean failure rate of 7.3% observed in the Spelling + Concepts group is not significantly different from the mean failure rate of 6.13% observed in the Spelling Only group ($t = -0.51$, $df = 469$, $p = 0.3046$). This shows that no significant change in the rate of search failures, either positive or negative, results from the addition of concept-based searching to the enhanced food search with spelling error correction.

6.2.3 Conclusions

When measured against the control group, the mean number of failed searches is reduced by more than 20% for the Spelling Only group and 18.9% for the Spelling + Concepts Group. Analysis of these results shows a statistically significant reduction in the rate of searches failing to return an acceptable selection for both enhanced food search versions. No significant difference in failure rate is observed between the enhanced food search with spelling error

correction and the enhanced food search with both spelling error correction and concept-based searching.

Figure 6.2 illustrates the improvements in search failure rates over control with both groups.

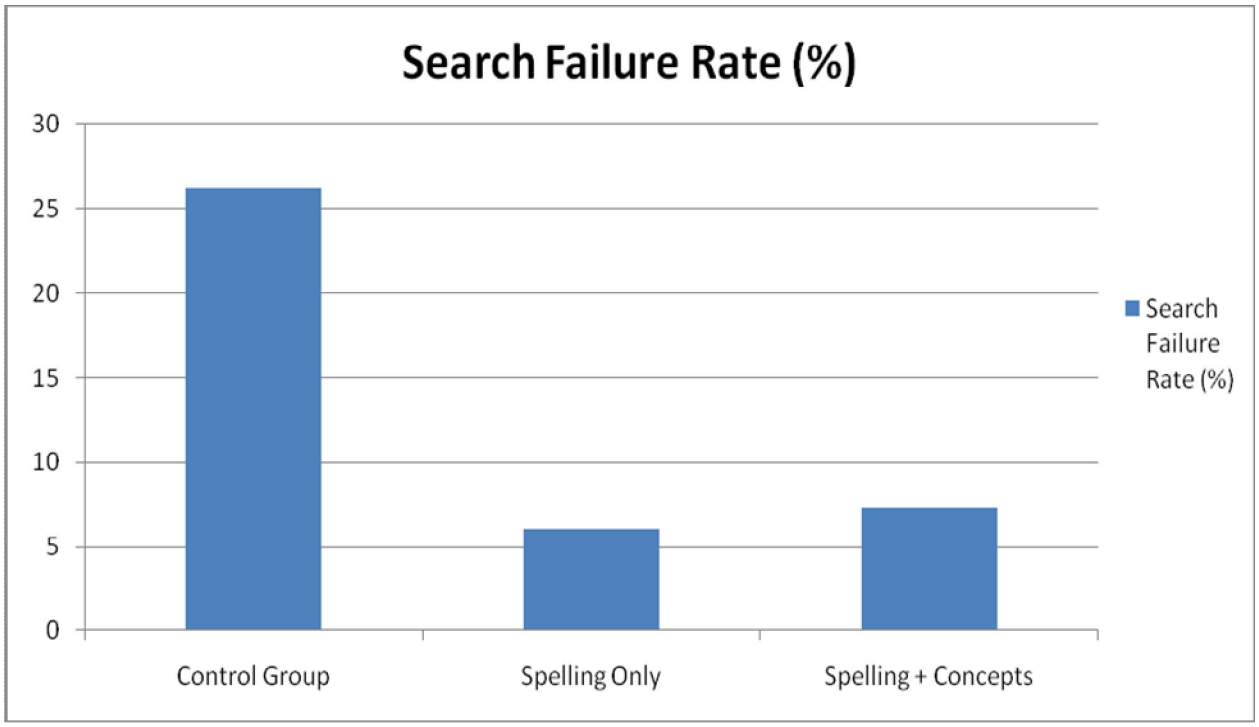

Figure 6.2 Mean Search Failure Rate (%)

6.3 H_A 3: No Significant Negative Impact on Precision

"Subjects using the enhanced search models will not receive significantly more search results than subjects using the current DINE Healthy food search."

Precision is a measure of the fraction of returned results that are relevant to the information need. It is usually expressed as the quotient produced by dividing the number of relevant search results by the total number of results returned. Because relevant result counts in this domain are constant across all groups and relatively small, the mean number of search results returned on text-based queries is used as a convenient approximation of precision.

6.3.1 The Spelling Only Group

Table 6.8 shows the results of a Two-Sample t-Test comparing the mean search results per query for the Control and Spelling Only groups. N is equal to the number of text-based queries conducted in each group.

Group	N	Mean (results)	STDV
Control	457	108.42	730.3
Spelling Only	281	70.75	100.5
Test	**DF**	**T**	**P**
Two-Sample t-Test	484	1.09	0.1390

Table 6.8 t-Test of Mean Results per Query for Spelling Only Group

Data collected from the Control group consists of 457 observations with a mean result per query count of 108.42. The Spelling Only group data consists of 281 observations with a mean result per query count of 70.75. The difference between mean results per query is not statistically significant ($t = 1.09$, $df = 484$, $p = 0.1390$). The rejection of this null hypothesis demonstrates that significantly more search results are not returned by the enhanced food search with spelling error correction. Further, the mean number of results generated per query is actually slightly less than with the current DINE Healthy food search, although not to a statistically significant degree.

6.3.2 The Spelling + Concepts Group

Table 6.9 shows the results of a Two-Sample t-Test comparing the mean search results per query for the Control and Spelling + Concepts groups.

Group	N	Mean (results)	STDV
Control	457	108.42	730.3
Spelling + Concepts	327	115.06	481.6
Test	**DF**	**T**	**P**
Two-Sample t-Test	777	-0.15	0.4391

Table 6.9 t-Test of Mean Results per Query for Spelling + Concepts Group

Data collected from the Spelling + Concepts group consists of 327 observations with a mean result count of 115.06. The difference between mean number of results per query for the Control group and Spelling + Concepts group is not statistically significant ($t = -0.15$, $df = 777$, $p = 0.4391$). Although the mean result count does exceed the Control group mean by 6.64 results, this increase fails to meet the standard test for statistical significance of $p < 0.05$.

Table 6.10 shows the results of a Two-Sample t-Test comparing the mean search results per query for the Spelling Only and Spelling + Concepts groups.

Group	N	Mean (results)	STDV
Spelling Only	281	70.75	100.5
Spelling + Concepts	327	115.06	481.6
Test	**DF**	**T**	**P**
Two-Sample t-Test	359	-1.62	0.0527

Table 6.10 t-Test of Mean Results per Query for Spelling Only and Spelling + Concepts

This test shows that the mean of 115.06 results per query observed in the Spelling + Concepts group is not significantly different from the mean of 70.75 results observed in the

Spelling Only group (df = 359, t = -1.62, p = 0.0527). Although by a narrow margin, this indicates that the addition concept-based searching to the enhanced food search with spelling error correction does not decrease precision by a statistically significant amount.

6.3.3 Conclusions

In this experiment, the mean number of search results generated per text-based query is observed to decrease by 37.7 results for subjects using the enhanced food search with spelling error correction. Although this difference is not significant (i.e. p = 0.1390), on the surface it is still somewhat surprising. This is because a slight decrease in precision is a typical when modifying a search system with the intent of increasing recall. That being said, no change of statistical significance is detected in either direction.

The mean result count for the Spelling + Concepts group is observed to rise by 6.6 results compared to the Control group, but his also fails to indicate a statistically significant change in mean results per query. In addition, the Spelling Only and Spelling + Concepts means are not significantly different. These findings demonstrate that subjects using the enhanced food search versions do not receive more search results per query than subject using the current DINE Healthy food search.

Figure 6.3 illustrates the similarity of mean result counts for both experimental groups and the Control group.

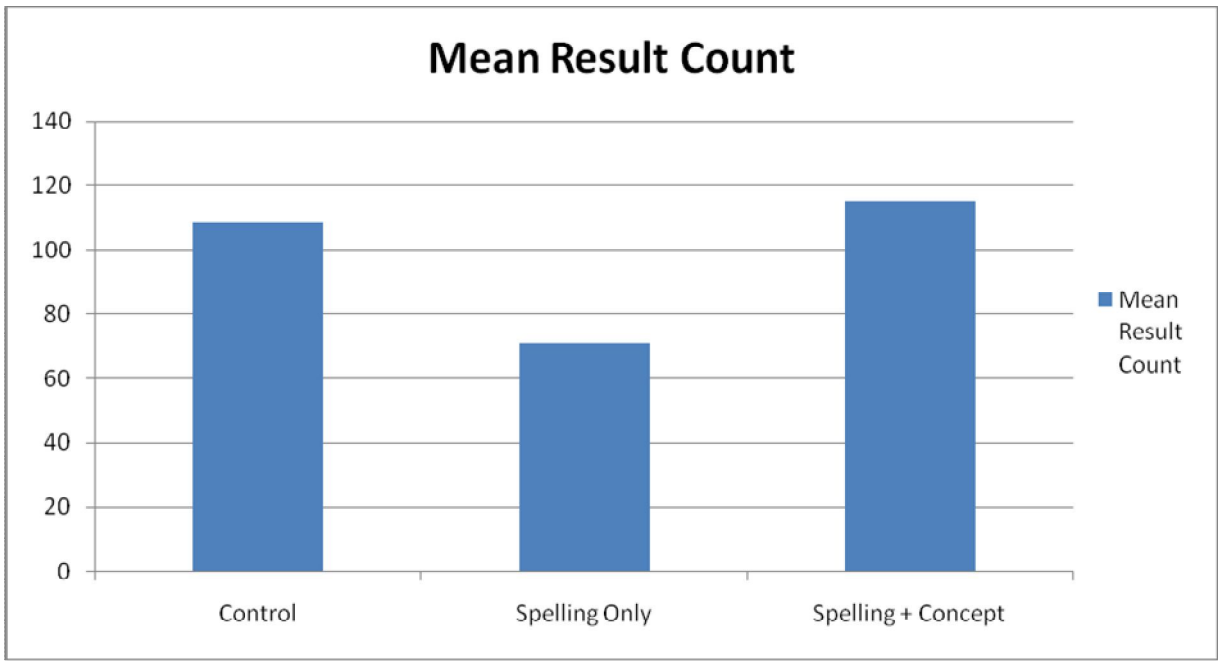

Figure 6.3 Mean Search Results per Text-Based Query

6.4 H_A 4: Reduction of Search Refinements

"Subjects using the enhanced search models will require fewer search refinements to locate food items than subjects using the current DINE Healthy food search."

When measuring the number of search refinements, several things must be considered:

- Some attempts to locate a single food item span multiple searches.
- Not every food search ends in a food item selection.
- Search refinements can consist of both text-based queries and category filtering.

For these reasons, evaluation of this hypothesis is based on the number of search refinements (text-based and filtering) required per food item selection.

6.4.1 The Spelling Only Group

Data collected from the Control group consists of 218 observations with a mean query refinement count per food item selection of 2.69. The Spelling Only group data consists of 200

observations with a mean of 2.23 query refinements per food item selection. The difference between means for the Control group and Spelling only group is statistically significant ($t = 2.01$, $df = 373$, $p < 0.03$). This demonstrates that subjects using the enhanced food search with spelling error correction require fewer query refinements to locate food items than subjects using the current DINE Healthy food search.

Table 6.11 shows the results of a Two-Sample t-Test comparing the mean search refinements per food item selection for the Control and Spelling Only groups. N is equal to the total number of food item selections for each group.

Group	**N**	**Mean**	**STDV**
Control	218	2.69	2.8
Spelling Only	200	2.23	1.8
Test	**DF**	**T**	**P**
Two-Sample t-Test	373	2.01	0.0228

Table 6.11 t-Test of Mean Search Refinements for Spelling Only Group

6.4.2 The Spelling + Concepts Group

Data collected from the experimental Spelling + Concepts group consists of 260 observations with a mean count of 1.42 refinements per query. The difference between mean search refinements per query is statistically significant ($t = 6.33$, $df = 259$, $p < 0.0001$), which demonstrates that subjects using the enhanced food search version with both spelling error correction and concept-based searching require fewer query refinements to locate food items than subjects using the current DINE Healthy food search.

Table 6.12 shows the results of a Two-Sample t-Test comparing the mean search refinements per food item selection for the Control and Spelling + Concepts groups.

Group	N	Mean	STDV
Control	218	2.69	2.8
Spelling + Concepts	260	1.42	1.0
Test	**DF**	**T**	**P**
Two-Sample t-Test	259	6.33	< 0.0001

Table 6.12 t-Test of Mean Search Refinements for Spelling + Concepts Group

A Two-Sample t-Test shows that the mean of 1.42 refinements for the Spelling + Concepts group is significantly less than the mean of 2.23 refinements for the Spelling Only group ($df = 284$, $t = 5.71$, $p < 0.0001$). This demonstrates that the addition of concept-based searching to the enhanced food search with spelling error correction results in significantly fewer search refinements per food item selection. Table 6.13 shows the results of a Two-Sample t-Test comparing the mean search refinements per food item selection for the Spelling Only and Spelling + Concepts groups.

Group	N	Mean	STDV
Spelling Only	200	2.23	1.8
Spelling + Concepts	260	1.42	1.0
Test	**DF**	**T**	**P**
Two-Sample t-Test	284	5.71	< 0.0001

Table 6.13 t-Test of Mean Search Refinements for Spelling Only and Spelling + Concepts

6.4.3 Conclusions

Analysis of experimental data shows that mean query refinements per food item selected are reduced by 17.1% for the Spelling Only group and 47.2% for the Spelling + Concepts group. These results confirm the hypothesis that a reduction in mean number of query refinements can be achieved using the enhanced food search versions. Futher, a significant reduction in mean query refinements is achieved by the addition of concept-based searching to the enhanced food search with spelling error correction.

Figure 6.4 illustrates the reduction in means of both experimental groups compared to the control group.

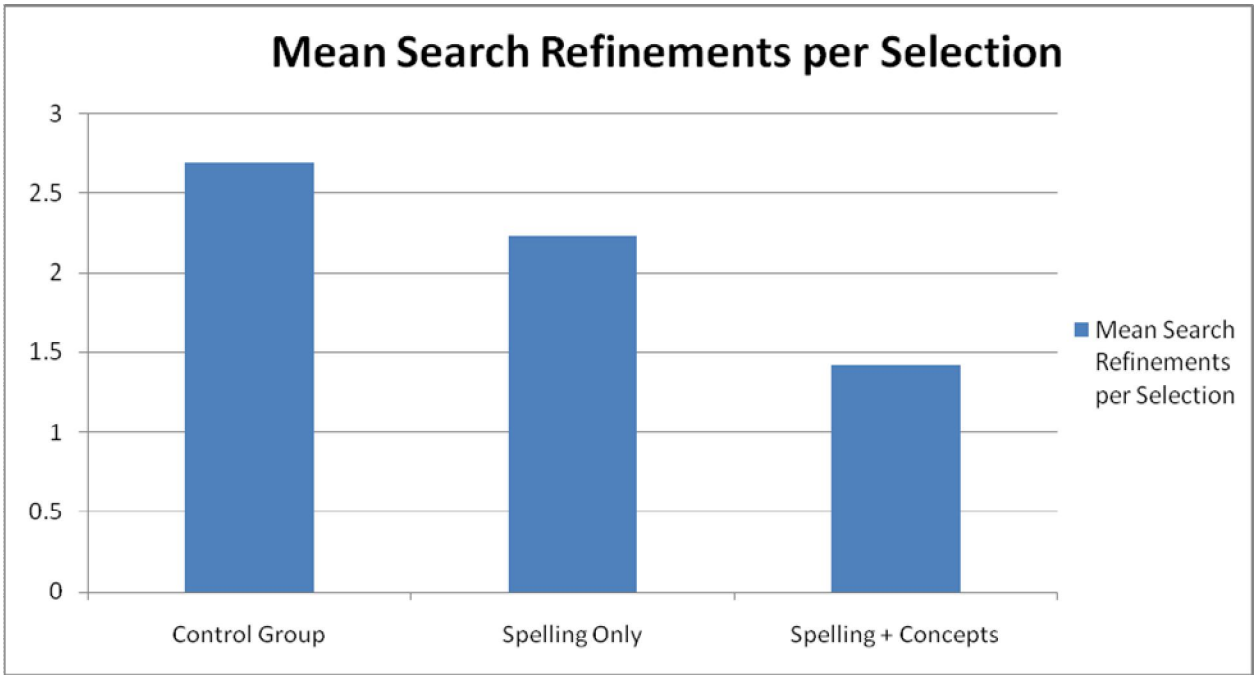

Figure 6.4 Mean Search Refinements per Selection

CHAPTER 7: Summary and Future Work

7.1 Spelling Error Correction

The implementation of spelling error correction techniques makes an obviously positive contribution to the DINE Healthy food search system. Statistically significant reductions in mean search time, search failure rates, and required query refinements are all demonstrated by the experiments. However, more insight can be gained from examining the rate of occurrence of spelling errors within the domain and the effectiveness of the spelling error correction system at resolving these errors.

7.1.1 Establishing Need

An analysis of 1,070 text-based queries entered by subjects in all three groups (i.e. Control, Spelling Only, and Spelling + Concepts) reveals that 368 contain one or more spelling errors. The total number of errors in all queries is 374, with only 6 queries containing more than one error. In the Control group, this results in 38.5% of all text-based queries failing to return the desired results due to spelling errors. The rate of occurrence of spelling errors across all groups is 34.4%, which actually exceeds the rate of over 26% stated to occur in web search engines (Wilbur et al. 2006). From this data, it is apparent that sufficient spelling errors occur in user queries to negatively affect the food search system. Table 7.1 shows the breakdown of spelling errors by group.

	Selections	Queries	Errors	% Errors
Control Group	218	457	176	38.5
Spelling Only	200	285	87	30.5
Spelling + Concepts	260	328	105	32.0
All Groups	678	1070	368	34.4

Table 7.1 Spelling Errors in Text-Based Queries

7.1.2 Specific Contribution

Further, testing of the automatic spelling error correction system reveals that 91% of these misspellings are automatically corrected to the intended word without user intervention. Interestingly, this figure would be more than 96% if not for numerous instances in which the misspelling “cantelope” is replaced with “antelope” rather than the desired “cantaloupe” (More discussion of this phenomenon is provided in section 7.1.3).

With the knowledge that 91% of the automatic corrections result in effective treatment of spelling errors, the impact on overall search performance is obvious. More than 38% of text-based queries are affected by spelling errors in the Control group, but this drops to a rate of less than 3% for each of the experimental groups.

Table 7.2 compares the percentage of queries containing spelling errors with the percentage of queries affected by those errors due to automatic spelling error correction.

Group	% Spelling Errors	% Queries Affected
Control	38.5	38.5
Spelling Only	30.5	2.7
Spelling + Concepts	32.0	2.9

Table 7.2 Effects of Spelling Error Correction on Text-Based Queries

7.1.3 Lessons Learned

Some additional issues discovered during the course of this research might be of interest to others attempting similar experiments with spelling error correction. One of particular interest has been discovered since the completion of data collection and analysis. Although performance of the spelling error correction system is good (i.e. 91% of auto-corrections are effective), this figure could be improved to over 96% by dealing with a single, common misspelling.

The misspelling of “cantaloupe” most often observed in this experiment is “cantelope”. This presents a problem for the spelling correction system because the most similar word in the dictionary is “antelope” (i.e. only 1 transformation), and the problem is compounded because the terms “cantaloupe” and “antelope” occur with similar frequency in the corpus. The result is that automatic corrections for “cantaloupe” are often unsuccessful. One potential solution to this problem involves exploiting the concept-based searching system to handle common misspellings by adding a relationship associating them.

Another, spelling error correction issue involves the size of the corpus from which the dictionary of terms is created. The corpus used in this experiment consists of only 3,700 unique terms, so automatic correction of spelling errors is extremely effective in most cases. However, with domains requiring a much larger corpus, the same benefits may not be realized. In such a situation, a separation of the detection and correction of spelling errors involving user interaction may be required.

7.2 Concept-Based Searching

The enhanced food search with both spelling error correction and concept based searching offers obvious improvements over the current DINE Healthy food search. Significant improvements are also shown to result from the addition of concept-based searching to the enhanced food search with spelling error correction. A reduction in mean search time and mean number of query refinements is demonstrated without significant decrease in precision or increase in the rate of search failures.

7.2.1 Query Expansion

With the knowledge that concept-based searching techniques are valuable to the food search, the next challenge becomes isolating the source of this value. In furtherance of this goal, the 687

text-based queries from all groups that immediately precede food item selections are examined. The selected food item IDs and query texts are processed by a food search simulation algorithm that measures the position of the selected food items in each search result set.

A separate simulation is run on all queries for both enhanced food search versions. This accomplishes two things:

1 Instances are identified in which concept-based searching enables the selection of a food item that would not otherwise be possible with the given query.
2 Comparisons can be drawn between the sorting capabilities of both enhanced food search versions on a per-query basis.

The implication of the first statement is that for each such occurrence, the use of concept-based searching techniques has prevented the need for additional searching and thereby reduced search time. Analysis reveals that in 27 cases (10.4% of food item selections from the Spelling + Concepts group); the food item selection would have required additional user action without the use of concept-based searching.

Table 7.3 summarizes these findings.

Group	N	Found w/ Spelling Only	Found w/ Spelling + Concepts
Spelling + Concepts	260	233*	260

*27 of 260 food items selected by subjects from the Spelling + Concepts group would not have been possible without concept-based searching

Table 7.3 Food Item Selections Absent without Concept-Based Searching

7.2.2 Sorting Results by Relevance

The second issue explored in these simulations is the position of food items in the result set generated by each enhanced food search version. A Two-Sample t-Test is used to evaluate the mean order of occurrence of all selected food items in search result sets generated by both versions. Obviously, the 27 queries that are only completed through concept-based searching cannot be compared, and are eliminated from mean calculations.

The test reveals that with concept-based searching, the mean position of selected food items is on average in position 12 in the result set. With spelling error correction only, the mean position is on average 32. The difference between means is significant ($t = 9.61$, $df = 650$, $p < 0.0001$), which indicates that the enhanced food search with both spelling error correction and concept-based searching is more effective at ordering results than the food search with spelling error correction only.

Version Used to Process Queries	**N**	**Mean**	**STDV**
Spelling Only	651	31.89	62.44
Spelling + Concepts	651	11.9	30.25
Test	**DF**	**T**	**P**
Two-Sample Paired t-Test	650	9.61	< 0.0001

Table 7.4 t-Test of Mean Search Result Position

7.2.3 Lessons Learned

During the course of this research, lessons have been learned that may be of importance to others pursuing similar projects. An issue of particular importance relates to the integration of different types of advanced searching techniques into a single system, particularly within a

specific, limited domain. In such cases, special care must be taken to ensure that the normal behavior of one system is not adversely affected by the behavior of others.

For example, during early development of concept-based searching algorithms, strange results were detected for some concepts (i.e. certain parts of the ontology were unreachable by query). The effect was caused by the spelling error correction system altering concepts before they reached the ontology for processing. This resulted from the inclusion of domain-specific terminology in the ontology that was not present in the dictionary of terms. In essence, concepts missing from the dictionary were automatically identified as non-words and "corrected". In this research, the issue is resolved by ensuring that all ontological concepts are also present in the dictionary.

Additionally, the domain of food item description information has the simplifying characteristic of a basically hierarchical structure. For example, food items can be classified into more and more specific categories with little difficulty. One interesting exception arises from the fact that some ambiguity exists as to the classification of certain items as either fruits of vegetables. This issue is resolved by taking advantage of the flexible nature of ontologies and classifying such items as both fruits and vegetables. Such problems may be difficult to solve in a more complex domain.

7.2.4 Future Work

A potential area for future work exists in identifying methods for developing more mature weighting schemes to use with query expansion and relevance ranking algorithms. The proximity weighting function is of particular interest, because it would most like vary tremendously across different domains. This is because the function used in this research attempts to take advantage of the known structure of DINE Healthy food item descriptions. This

scenario may be impossible or at least much different in other domains, specifically where text proximity is less related to relevance.

Potential for future work also exists with integrating popularity, or observed food item selection frequencies, into the relevance weighting algorithm. This could take the form of capturing *a priori* knowledge from large-scale search log analysis, or it could consist of an adaptive model designed to learn from user selections over time. An adaptive model appears particularly interesting because of the potential for automatically suppressing outdated or regionally unpopular food items. In either case, it seems intuitive that knowledge about food item selection frequency would be beneficial as a mechanism for adjusting relevance weights.

7.3 Conclusion

The application of advanced searching techniques to the DINE Healthy food search provides measurable, statistically significant benefits. This assertion is supported by the analysis of data collected from human subjects using the current DINE Healthy food search version as a control, and two experimental groups using enhanced food search versions.

A reduction in mean search time per food item selection of 52% is demonstrated for the enhanced food search with both spelling error correction and concept-based searching. Further, a deeper analysis of mean search times demonstrates that a statistically significant reduction in mean search time can be attributed to the application of concept-based searching techniques. These benefits are achievable without a significant decrease in precision.

REFERENCES

Belkin, Nicholas J. Some(what) Grand Challenges for Information Retrieval. *ACM SIGIR Forum.* Volume 42, Issue 1. Pages 47-54. June, 2008.

Clark, Peter, John Thompson, Heather Holmback, Lisbeth Duncan. Exploiting a Thesaurus-Based Semantic Net for Knowledge-Based Search. *Proc 12th Conf on Innovative Applications of AI.* 2000. (p.988 – 995)

GSC: Genomic Standards Consortium. *Food Ontology Project.* Updated November 26, 2008. December 5th, 2008. http://gensc.org/gc_wiki/index.php/Food_Ontology_Project

HEI: Healthy Eating Index. *United States Department of Agriculture Center for Nutrition Policy and Promotion.* Updated December 04, 2008, January 22nd, 2009. http://www.cnpp.usda.gov/HealthyEatingIndex.htm

Hull, David. Using Statistical Testing in the Evaluation of Retrieval Experiments. *Proceedings of the 16th annual international ACM SIGIR conference on Research and development in information retrieval.* 1993. (p.329 – 338)

Ide, Nocholas C., Russell F. Loane, Dina Demner-Fushman. Essie: A Concept-based Search Engine for Structured Biomedical Text. *Journal of the American Medical Informatics Association*. Volume 14, Number 3. 2007.

Jasco, Peter. Query refinement by word proximity and position. *Online Information Review.* Volume 28. 2004. (p.158-161)

Jurafsky, Daniel and James H. Martin. *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition.* Upper Saddle River, NJ: Prentice-Hall, Inc. 2000.

Kukich, Karen. Technique for Automatically Correcting Words in Text. *ACM Computing Surveys (C SUR).* Volume 24, Issue 4. Pages 377-439. December, 1992.

Manning, Christopher, Prabhakar Raghavan and Hinrich Schütze. *Introduction to Information Retrieval.* New York, NY: Cambridge University Press. 2008.

Noy, Natalya F. and Deborah L. McGuinness. Ontology Development 101: A Guide to Creating Your Fist Ontology. *Stanford Knowledge Systems Laboratory Technical Report* KSL-01-05. March 2001.

Probst, Yasmine C. and Linda C. Tapsell. Overview of Computerized Dietary Assessment Programs for Research and Practice in Nutrition Education. *Journal of Nutrition Education and Behavior.* Volume 37. 2005. (p.20 – 26)

Pyrczak, Fred. *Success at Statistics (3rd Edition).* Glendale, CA: Pyrczak Publishing. 2004.

Salton, Gerard and Christopher Buckley. Term-Weighting Approaches in Automatic Text Retrieval. *Information Processing & Management*. Volume 24. 1988. (p.513 – 523)

Search Tools Reports: Searching for Text Information in Databases. *Search Tools Consulting.* Updated 2003-07-23. 26 Nov. 2008 http://www.searchtools.com/info/database-search.html

Silverstein, Craig et al. Analysis of a Very Large Web Search Engine Query Log. *ACM SIGIR Forum*. 1999. 1: (p.6 – 12)

TREC: Text REtrieval Conference. *National Institute of Standards and Technology.* Updated August 28, 2008. http://trec.nist.gov/overview.html

Tunkelang, Daniel. Resolving the Battle Royale between Information Retrieval and Information Science. *Information Seeking Support Systems Workshop*. Chapel Hill, NC, USA. 2008.

Voorhees, Ellen M. TREC: Improving Information Access through Evaluation. *Bulletin of the American Society for Information Science and Technology.* Volume 32, No. 1. October/November 2995

Wilbur, W. John, Won Kim, and Natalie Xie. Spelling Correction in the Pubmed Search Engine. *Information Retrieval Boston.* November 2005. (p.543 – 564)

APPENDIX A

Food Item Selections Made by Human Subjects:

Food Id	Unique food item identifier in the DINE Healthy Food Database
Description	Text description of the food item
N	Number of times selected by human subjects
T	C = Correct, A = Acceptable, I = Incorrect

1. Mashed Potatoes

Food Id	Description	N	T
007212	POTATOES, Mashed, Boston Market	31	C
007215	POTATOES, Mashed, granules w/milk, prepared w/water & fat	1	C
007217	POTATOES, Mashed, home prepared w/whole milk & salt	1	C
007218	POTATOES, Mashed, home prepared w/whole milk, margarine & salt	1	C

2. Cantaloupe

Food Id	Description	N	T
005893	MELON, Cantaloupe, raw, cubed	30	C
005892	MELON, Cantaloupe, raw (5 in diameter)	2	C
009666	WENDY'S, Cantaloupe, fresh, sliced	1	A
000063	ANTELOPE, Roasted	1	I

3. Tomato

Food Id	Description	N	T
009266	TOMATOES, Raw (2-3/5 in diameter)	22	C
009267	TOMATOES, Raw (6.6 cm diameter) (Canada)	3	C
009274	TOMATOES, Red, ripe, raw, June thru October average	3	C
009275	TOMATOES, Red, ripe, raw, November thru May average	3	C

009270	TOMATOES, Raw, June to October (Canada)	1	C
009264	TOMATOES, Green, raw (2-3/5 in diameter)	1	I
009258	TOMATOES, Canned, stewed, Original Style, Del Monte	1	C

4. Scrambled Eggs

Food Id	Description	N	T
003886	EGGS, Scrambled w/margarine & whole milk	28	C
003887	EGGS, Scrambled, fast food breakfast	6	C

5. Strawberries

Food Id	Description	N	T
008840	STRAWBERRIES, Raw	32	C
008838	STRAWBERRIES, Frozen, sweetened, whole	1	A
004778	GUAVAS, Strawberry, raw	1	I

6. Croissant

Food Id	Description	N	T
003616	CROISSANTS, Plain, prepared w/butter (4-1/2 in x 4 in x 1-3/4 in)	14	C
003615	CROISSANTS, Plain, Dunkin' Donuts	6	C
003604	CROISSANTS, All butter & pre-sliced all butter, L'Original, Sara Lee	3	C
003605	CROISSANTS, All butter, petite size, L'original, Sara Lee	3	C
003608	CROISSANTS, Cheese, prepared w/butter (4-1/2 x 4 x 1-3/4 in)	1	C
007771	ROY ROGERS, Crescent roll	3	A
007715	ROLL, Dinner, commercial	2	A
003611	CROISSANTS, Egg & cheese, fast food breakfast	1	I

7. Broccoli

Food Id	Description	N	T
001504	BROCCOLI, Raw, chopped	7	C
001505	BROCCOLI, Raw, cooked, boiled, drained, chopped	7	C
001491	BROCCOLI, Boiled, drained (Canada)	6	A
001494	BROCCOLI, Flower clusters, raw	5	C
001490	BROCCOLI, Boiled with salt, drained (Canada)	3	A
001508	BROCCOLI, Stalks, raw	3	A
001493	BROCCOLI, Cooked, boiled, drained, with salt	2	C
001506	BROCCOLI, Raw, spears, cooked, boiled, drained (5 in spear)	1	A

8. Raisins

Food Id	Description	N	T
007477	RAISINS, Golden seedless, not packed	23	C
007478	RAISINS, Golden seedless, packed	2	C
005997	MIXED FRUIT, Dried, fruit bits & raisins, Sun-Maid	2	I
007480	RAISINS, Seedless, not packed	1	C
005999	MIXED FRUIT, Dried, prunes, apricots, apples & pears	1	I
007301	PRUNES, Dehydrated, uncooked	1	I
000106	APRICOTS, Dried, halves, Sunsweet	1	I
004714	GRAPES, Canned, Thompson seedless, water pack	1	I
000037	ALMONDS, Dried, blanched	1	I

9. Spaghettios & Meatballs

Food Id	Description	N	T
008748	SPAGHETTIO'S, Pasta w/meatballs in tomato sauce, Franco-American	16	C
008699	SPAGHETTI & MEATBALLS, In tomato sauce, Chef Boy-ar-dee	6	A

008747	SPAGHETTIO'S, In tomato & cheese sauce, Franco-American	5	C
000006	ABC'S & 123'S, W/mini meatballs, in tomato sauce, Chef Boy-ar-dee	2	A
000005	ABC'S & 123'S, W/mini meatballs in tomato sauce, microwave bowl, Chef Boy-ar-dee	2	A
006222	O-RINGS, W/mini meatballs in tomato sauce, bite size, Sir Chompsalot	1	A
008701	SPAGHETTI & MEATBALLS, In tomato sauce, microwave bowl, Chef Boy-ar-dee	1	A

10. Creamed Corn

Food Id	Description	N	T
003289	CORN, Yellow, canned, cream style	16	C
003279	CORN, Sweet, canned, cream style (Canada)	6	A
003285	CORN, White or yellow, cut off cob, cooked, boiled, drained	3	A
003284	CORN, White or yellow, canned, solids & liquid	2	A
003291	CORN, Yellow, canned, cream style, no salt added	2	C
000198	BABY CORN, Ka-me	2	I
003290	CORN, Yellow, canned, cream style, low sodium	1	C
000471	BABY FOOD, VEGETABLES, Creamed corn, 2nd Foods, Gerber	1	I
006050	MIXED VEGETABLES, Succotash, creamed corn & lima beans, canned	1	I

11. Pepperoni Pizza

Food Id	Description	N	T
006893	PIZZA, Pepperoni, Domino's	29	C
006873	PIZZA, Double cheese/pepperoni, Domino's	5	C

12. Banana

Food Id	Description	N	T
000611	BANANAS, Raw (8-3/4 inch long x 1-13/32 inch diameter)	33	C
004583	FRUIT SALAD, Fresh, apples, bananas, grapes, oranges & pears	1	I

13. McDonald's French Fries

Food Id	Description	N	T
005852	MCDONALD'S, Large french fries	20	C
005871	MCDONALD'S, Small french fries	9	C
005853	MCDONALD'S, Large french fries, unsalted	3	C
004072	FRENCH FRIES, Fried in vegetable oil, large, fast food side dish	2	A
004070	FRENCH FRIES, Fried in beef tallow & vegetable oil, regular, fast food side dish	1	A

14. Sliced Watermelon

Food Id	Description	N	T
009642	WATERMELON, Raw, sliced (10 in diameter x 1 in thick)	30	C
009641	WATERMELON, Raw, diced	3	A
009775	WENDY'S, Watermelon, fresh	1	A

15. Baby Carrots

Food Id	Description	N	T
001968	CARROTS, Baby, raw, medium sized (2-3/4 in long)	24	C
001975	CARROTS, Frozen, baby, whole, cooked	5	C
001969	CARROTS, Boiled, drained (Canada)	2	A
001976	CARROTS, Frozen, sliced, cooked, boiled, drained	1	C
001979	CARROTS, Raw (20 cm x 2.5 cm dia) (Canada)	1	A
001982	CARROTS, Raw, sliced, cooked, boiled, drained	1	C

16. Macaroni & Cheese

Food Id	Description	N	T
005650	MACARONI & CHEESE, Mix, prepared as directed, original, Kraft	19	C
005648	MACARONI & CHEESE, Mix, deluxe dinner, Kraft	15	C

17. Snickers Candy Bar

Food Id	Description	N	T
001842	CANDY BAR, Snickers, M&M Mars (2.07 oz bar)	30	C
001822	CANDY BAR, Milk chocolate w/peanuts (1.5 oz bar)	1	C
001798	CANDY BAR, 3 Musketeers, M&M Mars (2.13 oz bar)	1	A
001828	CANDY BAR, Milky Way Dark, M&M Mars (1.76 oz bar)	1	A
001844	CANDY BAR, Sweet chocolate (1.45 oz bar)	1	A

18. Budweiser

Food Id	Description	N	T
001163	BEER, Budweiser	33	C
009999	BEER, Light, Budweiser Select	1	I

19. Grape Soft Drink

Food Id	Description	N	T
004700	GRAPE SOFT DRINK	17	C
008303	SNAPPLE, Amazin' grape soda	12	A
005291	JUICE, Grape, canned or bottled	2	A
004697	GRAPE DRINK, W/vitamin C, canned (Canada)	1	C
005194	JUICE DRINK, Cranberry cocktail	1	I
009836	WILD BERRY, Soda, Health Valley	1	A

20. Deviled Eggs

Food Id	Description	N	T
003867	EGGS, Deviled, home prepared w/mayonnaise	32	C
003888	EGGS, Soft-boiled, large	1	A
003856	EGG WHITES, Raw, fresh & frozen	1	I