A COMPUTATIONAL APPROACH TO CONTEXTUAL INTERNET SEARCH USING MUTUAL INFORMATION WITH LEXICO-SYNTACTIC PATTERNS

By

Vinh X. Nghiem

A DISSERTATION PRESENTED TO THE GRADUATE SCHOOL OF THE UNIVERSITY OF FLORIDA IN PARTIAL FULFILLMENT OF THE REQUIREMENTS OF THE DEGREE OF DOCTOR OF PHILOSOPHY

UNIVERSITY OF FLORIDA

2009
ACKNOWLEDGMENTS

I thank my research advisor and supervisory committee chair, Dr. Paul Fishwick for his support, direction, and patience. I also thank my committee members: Dr. Doug Dankel, Dr. Jeff Ho, Dr. Manuel Bermudez, and Dr. Hana Filip for their advice and expertise. Lastly, and mostly, I thank my dear parents for their truly unconditional love.
TABLE OF CONTENTS

<table>
<thead>
<tr>
<th>SECTION</th>
<th>PAGE</th>
</tr>
</thead>
<tbody>
<tr>
<td>ACKNOWLEDGMENTS</td>
<td>3</td>
</tr>
<tr>
<td>LIST OF TABLES</td>
<td>6</td>
</tr>
<tr>
<td>LIST OF FIGURES</td>
<td>7</td>
</tr>
<tr>
<td>ABSTRACT</td>
<td>8</td>
</tr>
<tr>
<td>CHAPTER</td>
<td></td>
</tr>
<tr>
<td>1 INTRODUCTION</td>
<td>10</td>
</tr>
<tr>
<td>1.1 Motivation</td>
<td>10</td>
</tr>
<tr>
<td>1.2 Proposed Approach and Issues</td>
<td>11</td>
</tr>
<tr>
<td>1.3 Evaluation of Implementation</td>
<td>12</td>
</tr>
<tr>
<td>1.4 Contributions</td>
<td>12</td>
</tr>
<tr>
<td>1.5 Hypothesis</td>
<td>13</td>
</tr>
<tr>
<td>1.6 Outline of Dissertation</td>
<td>13</td>
</tr>
<tr>
<td>2 RELATED WORK</td>
<td>15</td>
</tr>
<tr>
<td>2.1 KeyConcept</td>
<td>15</td>
</tr>
<tr>
<td>2.2 Conceptual Graph Matching</td>
<td>16</td>
</tr>
<tr>
<td>2.3 Text Classification</td>
<td>17</td>
</tr>
<tr>
<td>2.4 Hierarchical Categorization of Web Pages</td>
<td>18</td>
</tr>
<tr>
<td>2.5 Contextual Search</td>
<td>20</td>
</tr>
<tr>
<td>3 BACKGROUND</td>
<td>22</td>
</tr>
<tr>
<td>3.1 Context</td>
<td>22</td>
</tr>
<tr>
<td>3.2 Mutual Information</td>
<td>22</td>
</tr>
<tr>
<td>3.3 Hearst Patterns</td>
<td>23</td>
</tr>
<tr>
<td>3.4 Search Engine Evaluation</td>
<td>24</td>
</tr>
<tr>
<td>3.5 Analysis of Variance</td>
<td>25</td>
</tr>
<tr>
<td>3.6 Spearman Rank Coefficient</td>
<td>27</td>
</tr>
<tr>
<td>4 IMPLEMENTATION</td>
<td>29</td>
</tr>
<tr>
<td>4.1 Final Information Score</td>
<td>29</td>
</tr>
<tr>
<td>4.2 Lexico-Syntactic Patterns</td>
<td>31</td>
</tr>
<tr>
<td>4.3 Algorithm</td>
<td>33</td>
</tr>
<tr>
<td>5 EXPERIMENT AND RESULTS</td>
<td>34</td>
</tr>
<tr>
<td>Table</td>
<td>Description</td>
</tr>
<tr>
<td>--------</td>
<td>-----------------------------------------------------------------------------</td>
</tr>
<tr>
<td>5-1</td>
<td>Comparison of results for the query “Vinh Nghiem in the context of person” (Vinh Nghiem mentioned as a person and not a pagoda, for example)</td>
</tr>
<tr>
<td>5-2</td>
<td>Comparison of results for the query “war in the context of religion” (mentions of war relating to religion)</td>
</tr>
<tr>
<td>5-3</td>
<td>Comparison of results for the query “mouse in the context of novel” (mentions of a mouse as a character in a novel)</td>
</tr>
<tr>
<td>5-4</td>
<td>Comparison of results for the query “Marlon Brando in the context of father”</td>
</tr>
<tr>
<td>5-5</td>
<td>Comparison of results for the query “Flash in the context of superhero and cartoon”</td>
</tr>
<tr>
<td>5-6</td>
<td>Comparison of results for the query “Magic Johnson in the context of basketball player (not coach)”</td>
</tr>
<tr>
<td>5-7</td>
<td>Comparison of results for the query “scandal in the context of politics”</td>
</tr>
<tr>
<td>5-8</td>
<td>Comparison of results for the query “love in the context of a song”</td>
</tr>
<tr>
<td>5-9</td>
<td>Comparison of results for the query “Beetle in the context of car”</td>
</tr>
<tr>
<td>5-10</td>
<td>Comparison of results matching the query “Michael Phelps in the context of student”</td>
</tr>
<tr>
<td>5-11</td>
<td>Comparison of results matching the query “death in the context of disease”</td>
</tr>
<tr>
<td>5-12</td>
<td>Final F Score values for all participants within each query</td>
</tr>
</tbody>
</table>
LIST OF FIGURES

Figure                                      page

4-1  Pseudocode for contextual search ................................................................. 33

5-1  Directions for sample query ........................................................................... 37

5-2  Spearman Correlation Coefficient of the correlation of the contextual search
engine result to the mean participant score for each query. A Spearman
coefficient closer to +1 and -1 represent positive and negative correlations, to
human judgment, respectively, while 0 represents no correlation ...................... 43

5-3  Comparison of average results for contextual search queries ....................... 43

A-1  Precision and recall values for different methods of calculating composite
mutual information score across multiple patterns ................................................. 52

A-2  Output for results matching the query “Vinh Nghiem in the context of person”... 52

A-3  Output for results matching the query “war in the context of religion” .......... 52

A-4  Output for results matching the query “mouse in the context of novel” .......... 54

A-5  Output for results matching the query “Marlon Brando in the context of father”.. 54

A-6  Output for results matching the query “Flash in the context of superhero and
cartoon” .............................................................................................................. 55

A-7  Output for results matching the query “Magic Johnson in the context of
basketball player (not coach)” .............................................................................. 56

A-8  Output for results matching the query “scandal in the context of politics” .... 57

A-9  Output for results matching the query “love in the context of a song” .......... 58

A-10 Output for results matching the query “Beetle in the context of car” .......... 59

A-11 Output for results matching the query “Michael Phelps in the context of
student” .............................................................................................................. 59

A-12 Output for results matching the query “death in the context of disease” ....... 60
The most popular current internet search engines today, including those of Google and Yahoo, rely on matching the exact key words from the user query with occurrences in the target web pages. The results returned by these current search engines to the user will not always match the query of the user. Even with the same words matched, there may be a difference in meaning or context. For example, “Vinh Nghiem” coincidentally is the name of a famous pagoda in Vietnam and also an author of this paper. Since the pagoda is so popular, search engines naturally rank web pages related to it higher than those of the person. Furthermore, a Google search for “Vinh Nghiem person” merely identifies all web pages that contain this string of three words, whereas an actual page of that person would not actually contain the word “person,” but rather a picture of the person and personal contact information. What is needed in these instances is a search engine that allows the user to specify that the query pertains to a context specified by the user, and only such web pages with instances of the query in the specified context should be offered to the user. The research produced an approach that allows existing web crawlers the means to index words into given contexts.
Here, I propose a novel approach for contextual search, using mutual information and lexico-syntactic patterns and I have performed a human subject study to test the efficacy of this approach. Results on contextual search queries with our subjects have been highly promising, with precision and recall rates both over 90%, compared to Google’s precision and recall rates of under 60%. The human subject experiment with sample queries shows consistently strong correlation (0.8-1.0) of the contextual search algorithm to human judgment of context. This research allows for contextual search on the internet, which is more accurate than existing means. It would also lead to insight on the effect the mutual information of words with lexico-syntactic patterns has on the determination of context in natural language processing.
CHAPTER 1
INTRODUCTION

1.1 Motivation

The most popular current internet search engines today, including those of Google and Yahoo, rely on matching the exact key words from the user query with occurrences in the target web pages. The results these current search engines return to the user will not always match the query the user intended. Even with the same words used when searching, there may be a difference in meaning or context.

For example, “Vinh Nghiem” coincidentally is the name of a famous pagoda in Vietnam and also the author of this proposal. A Google or Yahoo search for “Vinh Nghiem” is saturated with links about the pagoda. Since the pagoda is so popular, search engines naturally rank web pages related to it higher than those of the person. Furthermore, a Google search for “Vinh Nghiem person” merely identifies all web pages that contain this string of three words, whereas an actual page of that person would not actually contain the word “person,” but rather, for example, a picture of the person and personal contact information.

Other such examples abound. For example, the user may want to know about Marlon Brando, but not in the most popular context of actor, but rather as a father; in this case, the user would only want web pages returned that mentioned anything about Marlon Brando’s context as a father, such as the number of children he had, his philosophy as a parent, or how his children regarded him. Another example includes searching for “war in the context of religion,” to find references to war that are related to religion, but not necessarily with religion mentioned; i.e., “holy war.” Another practical example for conceptual search is “mouse in the context of a novel,” to find instances of a mouse mentioned in the context of a fictional story. A Google or Yahoo search will
instead give most results pertaining to novel genes within the mouse. Yet another example would be a query to find information about cartoons related to the superhero The Flash. A reasonable search query for Google might be “Flash superhero cartoons,” yet most results are related to web animation using Adobe Flash. Current search engine searches will simply give results containing the query words that are also the most popular or most cited pages, without any attention to context.

A search engine is needed that allows the user to specify that a query pertains to a context specified by the user, and only such web pages with instances of the query in the specified context should be offered to the user by the engine. I produced an approach that will allows existing web crawlers the means to index words into given contexts.

1.2 Proposed Approach and Issues

Here I propose a novel approach for contextual search, using mutual information and lexico-syntactic patterns. The approach decides whether or not a sentence encountered by a web crawler belongs in a desired context. For each word of the sentence, the algorithm finds how strongly that word in the sentence fits into specific patterns involving the context. For example, the query “Vinh Nghiem in the context of a person” may result in the algorithm encountering the sentence “Vinh Nghiem is a historic pagoda that attracts thousands of tourists annually,” to which the algorithm will apply each word to a pattern relating to the desired context. Examples of these resulting patterns are “person has a pagoda,” “person has tourists,” “person’s historic,” “person with tourists,” etc. One would not expect the examples above to occur frequently in the space of all written sentences. Conversely, the sentence containing “Vinh Nghiem Office : CSE 502” will result in the algorithm trying words from that
sentence with patterns involving the context; examples are “person’s office” and “person has an office” (along with “person has CSE” and person has “502”) and the first two patterns here should result in positive context determination. Details on pattern construction and context determination are described in Section 5.1 Description of Implementation.

1.3 Evaluation of Implementation

To assess quality of the algorithm on a large number of queries, a human study is used, where a fixed number of individuals tag sentences from all web pages used by the algorithm. The precision and recall for the output produced by the algorithm are calculated on the concept labels from the human study. How closely the algorithm performs relative to human judgment is measured using the Spearman rank coefficient on the mean participant scores for each web page of each query compared with the results produced by the algorithm for each.

1.4 Contributions

This research makes novel contributions to contextual search in the following ways:

It provides a mathematical and linguistic framework for measuring contextual similarity. Previous methods have used nonspecific tools such as vector representations or probabilistic methods. This is the first published result that uses a framework that both is mathematical as well as linguistic, as it considers the fundamental meaning of words to gauge contextual similarity.

It leads to a direct implementation for a contextual search engine. Dominant existing search engines currently match the query to words on a page rather than look for contextually similar words. The algorithm proposed in this research could thus be
used for internet searches when a user specifies a query with a context. Even if users do not have a need at that moment for a search within a context, existing search engines could still be augmented to find web pages with contextually similar words to free the user from having to create just the right search terms.

A rigorous empirical study was conducted to gauge the validity of the algorithm that arose from the research. Such a study was needed because existing internet search studies that have standard judged datasets do not focus on search within context. Moreover, Yahoo, which published the only other contextual search engine (but using standard probabilistic frameworks) used an in-house human subject study (Kraft et al. 2005).

Aside from direct contextual search applications, there are implied applications. For example, mutual information with specific lexico-syntactic patterns can be used in ontology creation, to use the plethora of data of the internet to automatically find fundamental relationships among words.

1.5 Hypothesis

Determination of a sentence into a given context can be done accurately, using mutual information and lexico-syntactic patterns. This can be applied to determine web pages that have sentences in context.

1.6 Outline of Dissertation

The document is organized as follows: In Chapter 2, I discuss prior related work, and in Chapter 3, the necessary background is provided. In Chapter 4, I discuss the algorithm and its implementation. In Chapter 5, I present experimental results, discuss the data and issues related to improving data results, as well as discuss issues related to implementing a practical contextual search system using the method. In Chapter 6, I
discuss issues for improvement and future directions. In Chapter 7, I present conclusions.
CHAPTER 2
RELATED WORK

Contextual internet search is concerned with identifying pages which contain a search term in a similar meaning to a target context. Work that has been done toward this end consists of both statistical methods and structural methods, each with its own limitations, which this work addresses.

2.1 KeyConcept

KeyConcept (Gauche 2004; Ravindrand 2004) is a conceptual search engine that indexes web pages based on words (like a traditional search engine), as well as on concepts. The system classifies documents into 1,564 concepts extracted from the Open Directory Project. The system was trained with 30 documents per concept, after determining that more training per concept would give marginal returns. Documents are represented with a vector-space method such that documents are vectors of predetermined words, while no mention was made of the dimensionality of such vectors. Documents are then classified by comparing the centroid of each concept to the incoming document vector. The retrieval results garnered a top-20 precision of 26.9%, which was not a great improvement over the top-20 precision of 22% using keyword-only searches.

This system has two weaknesses in philosophy alone: (1) documents are being represented as document vectors, or bags of words, without regard to how the words are structured relative to each other, much less the sense disambiguation of the word itself; i.e., is the appearance of bank in a document in reference to the mouth of a river or a monetary institution? The document vector simply indicates that some sense of the word is present somewhere in the document, and the other weakness is that (2) the classification is occurring for the entire web page as a whole. This precludes the
identification of any context local to a sentence but not sharing a global context. For example, a page that gives mainly information about a certain model car and then alludes to a famous person owning the same car may be globally classified under “cars” but a contextual search for the car under the context of a specified actor as the owner would miss the objective of the query completely. The weaknesses in philosophy reflect the weakness in performance.

2.2 Conceptual Graph Matching

Conceptual Graph Matching (Zhong 2002) has been used to solve a more limited problem. With the domain restricted to clothing sales by manually selecting such pages, descriptions of clothing items are represented using graphs. Such a repository is stored and matched with an incoming query. For example, the query “a garment with a pouch and sleeve” can be matched to “a shirt with a pocket and sleeve.” Similarity between graph nodes is gauged using the number of relation hops from WordNet.

Such an approach transcends the traditional statistical, words-in-a-bag approaches and swings to the other extreme of a highly structural representation. While this allows a much more robust and reliable similarity comparison, it should be noted that the limitations in the approach reflect the general difficulties of such a representation. The approach is limited in its assumptions that the vocabulary set is limited, that the word usage is limited to specific patterns (“a shirt with…”), and that semantic ambiguities are rare (a word only has one definition, and not many as in natural language). In addition, the approach using WordNet to ultimately gauge similarities between words is a severe limitation. WordNet is effective as a repository for all senses (definitions) of a word, and while it also contain semantic relations (hypernyms, hyponyms, etc.), it is egregiously incomplete. For example, “student’,
“researcher,” or “teacher” should all be hyponyms of “person,” but they are not in WordNet.

These limitations would preclude using such an approach to locate context on documents with diverse sentence structure and where word senses are highly ambiguous as in contextual search on the internet.

2.3 Text Classification

Text classification refers to the assignment of category labels to new documents. Machine learning methods abound with statistical techniques that can be used to model a new category based on a labeled set of training documents with known categories. The model can then be applied to a new document to learn their categories. A wide variety of text retrieval and machine learning techniques can be used to build category models (Dumais et al. 1998).

A classifier or a classifier-building algorithm cannot interpret text directly. Because of this, an indexing procedure that maps a text, \(d_j\), into a compact representation of its content needs to be uniformly applied to training, validation, and test documents. The choice of a representation for text depends on what one regards as the meaningful units of text (the problem of lexical semantics) and the meaningful natural language rules for the combination of these units (the problem of compositional semantics). Similarly to what happens in information retrieval, in text classification this latter problem is usually disregarded, and a text, \(d_j\), is usually represented as a vector of term weights \(d_j = \{w_{i,j}, \ldots, w_{|T|,j}\}\), where \(T\) is the set of terms (sometimes called features) that occur at least once in at least one document of \(T_r\), and \(0 \leq w_{k,j} \leq 1\) represents, loosely speaking, how much term \(t_k\) contributes to the semantics of document \(d_j\). Differences among approaches are accounted for by (1) different ways to understand what a term is and (2)
different ways to compute term weights. A typical choice for (1) is to identify terms with words. This is often called either the set of words or the bag of words approach to document representation.

It has been found in a number of experiments (Apté et al. 1994; Dumais et al. 1998; Lewis 1992; Merkl 1998), that representations more sophisticated than this do not yield significantly better effectiveness. In particular, some authors have used phrases, rather than individual words, as indexing terms (Fuhr et al. 1991; Schutze et al. 1995; Tzeras and Hartmann 1993), but the experimental results found to date have not been uniformly encouraging, most likely because although indexing languages based on phrases have superior semantic qualities, they have inferior statistical qualities with respect to word-only indexing languages: a phrase-only indexing language has “more terms, more synonymous or nearly synonymous terms, lower consistency of assignment (since synonymous terms are not assigned to the same documents), and lower document frequency for terms” (Lewis 1992). Although his remarks are about syntactically motivated phrases, they also apply to statistically motivated ones, although perhaps to a smaller degree.

2.4 Hierarchical Categorization of Web Pages

Text classification has garnered piqued interest for its possible application in automatically classifying Web pages under the hierarchical categories on popular Internet portals, such as Yahoo or the Open Directory Project. Classifying Web pages automatically has clear practical benefits, since the manual classification of a large enough subset of the Web is infeasible. With respect to previously discussed text classification applications, automatic Web page classification has two essential peculiarities: (1) The hypertextual nature of the documents: Links are a rich source of
information, as they may be understood as stating the relevance of the linked page to the linking page. Techniques exploiting this intuition in a text classification context have been presented (Attardi et al. 1998, Chakrabarti et al. 1998b, and Oh et al. 2000) and experimentally compared (Yang et al. 2002). (2) The hierarchical structure of the category set: This may be used, for example, by decomposing the classification problem into a number of smaller classification problems, each corresponding to a branching decision at an internal node. Techniques exploiting this intuition in a text classification context have been presented (Ferragina et al. 2005, Dumais and Chen 2000, Ruiz and Srinivasan 1999, Chakrabarti et al. 1998a, McCallum et al. 1998, and Koller and Sahami 1997).

The limitation of these techniques is that they globally classify the entire document and this precludes the identification of any context local to a sentence not sharing a global context. For example, a page that gives mainly information about a certain model car and then alludes to a famous person owning the same car may be globally classified under “cars” but a contextual search for the car under the context of a specified actor as the owner would miss the objective of the query completely. The KeyConcept conceptual search system has these same global context limitations, and it should not be surprising that KeyConcept is applying traditional text classification techniques. It is important to note that it would be infeasible to apply these text classification techniques locally to sentences, essentially treating the sentences like documents. Most are supervised learning techniques, meaning they have the requirement of manually labeling each training sentence, and the problem is that the training required to manually label every representative of positive and negative sentence for a given context would be enormous, especially given the vast number of
ways of expressing the same sentence. Even if the technique were unsupervised, only sentences similar in representational structure, rather than logical or semantic structure, are positioned near each other in the training space. Expressed another way, sentences that have the most number of words in common are grouped together and said to share the same context, but this is effectively the kind of string matching that search engines can already do without text classification.

2.5 Contextual Search

Most work in contextual search involves user profiles based on search history, and recent web navigation patterns. This background is used to refine future searches. Another area of learning focuses on context learning (Goker 1999, Belkin et al. 1982), which is based on relevant documents, query terms, document vectors, etc.

Nemrava (2006) proposed a method for categorizing the concepts of a proper noun using WordNet and then verifying them with lexical patterns on Google. His method was limited to only proper nouns and on the accuracy of Wordnet relationships.

Context as a query (Henzinger et al. 2003, Rhodes et al. 1996, Budzik et al. 1999, Czerwinski et al. 1999, Billsus, et. al 2005) treats the context as a background for a query search. Contextual ads (e.g., Google AdSense) explore this idea further. Finkelstein et al. (2002) describe IntelliZap, a contextual search system. IntelliZap defines context to be the text that is around the neighborhood of a query, rather than context that is specified by the user.

Kraft et al. (1995) propose the Yahoo Y!Q System, which does not impose this context limitation. The Y!Q system uses both statistical and linguistic metrics. In the Y!Q system, context is determined statistically by composing context vectors from each given context and the terms for the vectors are terms that co-occur frequently with the
concept word. The drawback of this system, which is addressed by the proposed method, is that only matching co-occurrences would result in many false positives; i.e., science frequently co-occurs with religion. High co-occurrences with specific lexico-syntactic patterns achieves more accurate context determination. Furthermore, the Y!Q system uses as its linguistic component the UMLS Metathesaurus which is a hand-coded hierarchy of medical terms. Context is then measured by the distance of two terms on the tree. Such a system is severely limited because (1) only a small, specialized subset is provided in the hierarchy and (2) the work to create and maintain such a system is impractical. Instead, the method proposed here analytically uses mutual information with linguistic information that is determined using the entire internet, which is self-maintaining.

Previous contextual internet search methods have severe limitations. Some use only statistical methods and do not account for fundamental semantic relationships between the target context and the words in question. Methods that do use structural methods, such as graph methods, suffer from strictly limited corpus domains, and, still, such corpuses must be maintained tediously by hand. The work in this dissertation does not have such limitations; semantic relationships are determined dynamically using patterns that denote semantic relationships, while using the entire internet instead of a limited corpus that is impractical to scale. The next chapter lays the background for the method used in this research, which transcends limitations of the previous approaches.
CHAPTER 3
BACKGROUND

3.1 Context

A formal definition of context has been proposed by Dey (2001) to be any information that can be used to characterize the situation of an entity, while an entity is a person, place, or object that is considered relevant to the interaction between a user and an application, including the user and applications themselves. This applies particularly nowadays in mobile computing where search results returned should be in the context of the user’s location, past computing history, etc. It is relevant to the work here, since entity is defined to be strictly a noun, and this is the assumption being made to the query and desired context.

Furthermore, Teevan (2004) introduces the concept of orienteering when users search, particularly on the internet. Orienteering refers to the nature of users to know the direction they want to explore in their search but not exactly what they are looking for. Orienteering further justifies the idea of searching by context.

3.2 Mutual Information

The mutual information of two words is a measure of how strongly one word influences the presence of another in the set of all documents, in our case, the internet (Clear 1997). For example, to determine how strongly the word piano influences the presence of recital, one assumes a null hypothesis that piano has no affect on the frequency of recital and from this assumption, anticipates how often piano recital is to be seen in the text by multiplying the instances of piano by the general probability of recital occurring. To determine how similar the actual observed joint frequency is to the null hypothesis frequency, a ratio is determined between the actual observed joint frequency of piano recital and the anticipated frequency with the null hypothesis.
assumed. A ratio greater than one implies that the null hypothesis is incorrect, that the presence of one word indeed increases the probability of the second word occurring. The larger the ratio is, the more the null hypothesis can be discredited. It follows that the mutual information, or MI, can be expressed as

\[
MI(piano, \text{recital}) = \frac{f("piano\ \text{recital}")}{f(piano) \cdot p(\text{recital})} = \frac{f("piano\ \text{recital}")}{f(piano) \cdot \frac{f(\text{recital})}{N}} = \frac{f(piano)}{f(piano) \cdot f(\text{recital})} \cdot N
\]

\[
MI(piano, \text{recital}) = \frac{f(piano)}{f(piano) \cdot p(\text{recital})} = \frac{f(piano)}{f(piano) \cdot \frac{f(\text{recital})}{N}} = \frac{f(piano)}{f(piano) \cdot f(\text{recital})} \cdot N \text{ where}
\]

MI(piano, recital) represents the mutual information of piano and recital, \( f(piano) \) represents the frequency or number of occurrences of piano, \( p(\text{recital}) \) represents the probability of recital, and \( N \) represents the total number of words in all documents. To generalize for words a and b:

\[
MI(a \text{ then } b) = \frac{f(a \text{ then } b)}{f(a) \cdot f(b)} \cdot N
\]

### 3.3 Hearst Patterns

Hearst patterns are lexico-syntactic patterns that indicate the existence of class/subclass relations in unstructured data source, such as in web pages. Examples of lexico-syntactic patterns that were described in Hearst (2004) are, where NP represents a noun phrase:

- \( NP_0 \text{ such as } NP_1, NP_2, \ldots, NP_{n-1} \text{ (and | or) } NP_n \)
- \( \text{such } NP_0 \text{ as } NP_1, NP_2, \ldots, NP_{n-1} \text{ (and | or) } NP_n \)
- \( NP_1, NP_2, \ldots, NP_{n-1} \text{ (and | or) other } NP_0 \)
- \( NP_0 \text{ (incl. —esp.) } NP_1, NP_2, \ldots, NP_{n-1} \text{ (and | or) } NP_n \)
- \( \text{and very common "}NP_i \text{ is a } NP_0\text{"} \)
Hearst noticed that from patterns above one can derive for all \( NP_i, 1 < i < n \), a hyponym \((NP_i, NP_0)\).

Here, specific patterns suggested by Hearst are attempted to be used for context determination. In addition, new patterns are used to try to capture fundamental semantic relationships and avoid simple correlation without context inclusion.

### 3.4 Search Engine Evaluation

In evaluating the quality of search engine results, traditionally two measures have been used: the precision and recall. Recall \((\rho)\) describes the fraction of correctly identified records retrieved, while precision \((\pi)\) describes the fraction of the retrieved records that are correctly identified. These measures are not just reciprocals of one another as a reading at first glance may imply. *Precision with respect to* \( c_i \) \((\pi_i)\) is precisely defined as the conditional probability \( P(\Phi(d_x, c_i) = T | \Phi(d_x, c_i) = T) \), that is, as the probability that if a random document \( d_x \) is classified under \( c_i \), this decision is correct. Analogously, *recall with respect to* \( c_i \) \((\rho_i)\) is defined as \( P(\Phi(d_x, c_i) = T | \Phi(d_x, c_i) = T) \), that is, as the probability that, if a random document \( d_x \) ought to be classified under \( c_i \), this classification is taken. To use jargon from logic, \( \pi \) may be viewed as the “degree of soundness” of the classifier with respect to \( C \), while \( \rho \) may be viewed as its “degree of completeness” with respect to \( C \). As defined here, \( \rho_i \) and \( \pi_i \) are to be understood as *subjective* probabilities, that is, as measuring the expectation of the user that the system will behave correctly when classifying an unseen document under \( c_i \). Here, \( FPi \) (*false positives with respect to* \( c_i \) , a.k.a. *errors of commission*) is the number of test documents incorrectly classified under \( c_i \); \( TNi \) (*true negatives with respect to* \( c_i \) ), \( TPi \) (*true positives with respect to* \( c_i \) ), and \( FNi \) (*false negatives with respect to* \( c_i \) , a.k.a. *errors of omission*) are defined accordingly.
Estimates (indicated by carets) of precision with respect to $c_i$ and recall wrt $c_i$ may thus be obtained as

$$\hat{\pi}_i = \frac{TP_i}{TP_i + FP_i}, \hat{p}_i = \frac{TP_i}{TP_i + FN_i}.$$  

However, it is very difficult, if not impossible, to directly apply these measurements to the evaluation of Web information retrieval systems due to the vast and dynamic nature of the Web. Two alternate measures have been proposed Vaughn (2004). The first is the quality of the result ranking, which is counterpart measure to precision. The quality of the result ranking is a correlation of search engine ranking and human ranking. The second alternate measurement, a counterpart measure to recall, gauges the ability of the search algorithm to retrieve top ranked pages as determined by human judgment. Any quality assessment used requires human judgment.

3.5 Analysis of Variance

Analysis of Variance, or ANOVA, is a technique for testing the hypothesis that sample means of several groups are derived from the same population (SABLE 1999). In this case, each group is a human participant to judge whether specific web pages are contextually relevant to a specified query. For the same query, if all the human participants have similar innate mechanisms for judging contextual relevance and are all engaging these mechanisms fully (i.e., they are fully participating), then there should not be much variation in the scores given to the web sites for that query. However, if the participants do differ appreciably, then there may be statistically significant differences in the participant scores for the web pages of each query. Data on the average web page relevancy scores for each query of each participant can be used to test the null hypothesis that the average relevancy scores for all web sites of a query does not vary significantly from participant to participant.
To decide whether or not the participants are producing the same judgments, we are concerned with the question of whether the differences (variance) in the relevancy scores for each query made by each participant are "large" compared to the differences (variance) in the means across all participants. In essence, an ANOVA computation reduces to comparing the variances among the means to the variances within the samples. What it takes to be "large enough" for the difference to be statistically significant depends on the sample sizes and the amount of certainty that we desire in our testing (that is, \( p \) values or levels of statistical significance that we typically use with all of our significance tests).

To evaluate whether the \( B_{SS} \) is large relative to the \( W_{SS} \), it is necessary to take into account the number of independent scores, or degrees of freedom (d.f.) that contribute to each of those sums. For the \( B_{SS} \), \( d.f. = k-1 \) where \( k \) is the number of comparison groups. In estimating the overall population mean from the set of sample means, one degree of freedom is lost because, once all but one of the sample means is known, assuming the size of each sample is also known, then the mean for the \( k \)th sample is fixed. For the \( W_{SS} \), \( d.f. = N - k \). One degree of freedom is lost in calculating the sample mean for each of the \( k \) samples, for a total of \( N - k \) degrees of freedom lost in calculations of the \( W_{SS} \). The degrees of freedom for \( T_{SS} \) is \( (k - 1) + (N - k) = N - 1 \).

We now define the F-ratio as

\[
\frac{B_{ss} / (k - 1)}{W_{ss} / (N - k)} = \frac{B_{MS}}{W_{MS}}
\]

Here, \( B_{MS} \) and \( W_{MS} \) refer to the "Between-Groups Mean Squares" and "Within-Groups Mean Squares," respectively. Each of these \( MS \) values is an estimate of the variances in what is conceived of as a total population from which the various samples
were drawn. \( B_{MS} \) is an estimate of the variance for the several sample means, and \( W_{MS} \) is an estimate of the variance of the scores from their respective sample means. If differences across sample means are "large" relative to differences within samples, then we should reject the null hypothesis that the samples are all drawn from the same population.

The decision of whether or not to reject the null hypothesis that the sample means are similar to each other requires that the value for \( F \) be compared with a critical value. And, in turn, just as for the t-test, the critical value differs with the number of degrees of freedom. Unlike the t-test, the critical value of \( F \) needed to reject the null hypothesis at any given level of significance (e.g. .05, .01, or .001) varies with two rather than only one indicator of degrees of freedom. The alpha level of \( F \) for rejecting the null hypothesis depends on both the Between and the Within groups degrees of freedom.

An additional difference between the \( t \) and \( F \) tests is that t-tests are commonly used in one-tailed tests that are *directional* (e.g. electricians' average income is higher than plumbers' average income) as well as in two-tailed tests that are *non-directional* tests (electricians' average income differs from plumbers' average income) (SABLE 1999). By contrast, in general, the F test is used in non-directional tests. That is, the alternative hypothesis is that \( \mu_1 \neq \mu_2 \neq \mu_3 \neq \mu_k \). The question addressed here is strictly whether group of participants differ from each other, and, if so, throw out the participants that are different until there is no statistical difference.

### 3.6 Spearman Rank Coefficient

Spearman's rank correlation provides a distribution-free test of independence between two variables. In this study, those two variables are the algorithm’s score of
contextual relevance for each query’s web page on a binary scale of \{0,1\} and the average human participant score on a scale of \{0,1\}.

Spearman’s rank correlation coefficient is calculated as:

\[
\rho = \frac{n(\sum x_i y_i) - (\sum x_i)(\sum y_i)}{\sqrt{n(\sum x_i^2) - (\sum x_i)^2} \sqrt{n(\sum y_i^2) - (\sum y_i)^2}}.
\]

\(\rho\) takes on values within the closed interval \([-1,1]\), where \((0,1]\) reflects positive correlation between the two variables (increasing the value of one variable increases the second), \([-1,0)\) reflects negative correlation (increasing the value of one variable decreases the second), and \(\rho=0\) indicates that the variables are independent.

In this study, the Spearman Rank Coefficient is used to quantify how strongly the algorithm matches human judgment in determining contextual relevance. A contextual search algorithm that agrees well with human judgment would result in a suitably practical application.

In the next chapter, these background concepts of mutual information and lexico-syntactic patterns are used to formulate the algorithm used in this work. The concepts of analysis of variance and spearman correlation coefficient will also be used in the next chapter in testing an implementation of the algorithm.
CHAPTER 4
IMPLEMENTATION

The concepts discussed in the background, namely, mutual information and lexico-syntactic patterns, are used to formulate a final information score which determines contextual relevance, given a context and words on a web page in the same sentence as a desired term. This final mutual information score is derived in Section 4.1. In Section 4.2 the patterns used is discussed. Also, throughout the chapter, issues involved in design decisions are discussed.

4.1 Final Information Score

The system proposed here tests each word in the sentence containing the keyword match (i.e., the query “Vinh Nghiem in the context of person” would flag all sentences containing “Vinh Nghiem” and test each word in the sentence to determine if the context exists with “person.” The determination for context of a word is not based on an unnatural distance in hyper-dimensional space, but on how likely it is to influence a fundamental semantic pattern, namely, through the mutual information of the word and the pattern.

In this research, the mutual information concept has been extended to include patterns, rather than just individual words. For example, to gauge how strongly the pattern person-has-an influences the word office, the mutual information formula is applied directly to the pattern, word pair:

$$MI("\text{person has an office}" ) = \frac{f("\text{person has an office}")}{f("\text{person has an}") \cdot f("\text{office}")} \cdot N$$

To generalize:
\[ MI(\text{context + pattern, word}) = \frac{f(\text{context + pattern + word})}{f(\text{context + pattern}) \cdot f(\text{word})} \cdot N \]

The proposed method uses several patterns, and in turn, mutual information scores from several patterns. The scores are combined into a total mutual information score, which reflects how many observed patterns there actually are in the data relative to how many such patterns are expected:

\[ \text{Total MI}(\text{context + pattern skeleton, word}) = \sum_{i=1}^{\text{number of patterns}} \frac{f(\text{context + pattern + word})}{\sum_{i=1}^{\text{number of patterns}} f(\text{context + pattern}) \cdot f(\text{word})} \cdot N \]

The total mutual information score can be high with many contexts and not just the desired context. For example, in the search, “war in the context of religion,” if a word such as “science” is encountered on a page, “science” may have a high total mutual information score with “religion,” but it is also high with “war,” and so it should be concluded that “science” is not especially contextually relevant to the desired context of “religion.” Thus, the final mutual information score is computed as follows:

\[ \text{Final MI(context, pattern, word)} = \frac{\text{Total MI(context, pattern, word)}}{\text{Total MI(query, pattern, word)}} \]

An issue arises from how to consolidate the individual point-wise mutual information scores from many patterns. Possibilities include averaging or taking the highest score. A theoretically elegant solution is to calculate the final mutual information score as discussed. Figure A-1 shows the experimental results for the precision and recall rates on the test data, compared to alternative calculations, which confirms this intuition.
4.2 Lexico-Syntactic Patterns

The method is contingent on the choice of pre-constructed patterns, which reflect a fundamental relationship to the context and which best fit the training data. The resulting patterns are then applied on unseen test data. For a given context and for each word on a web page, the method queries the Bing Search Engine API (due to no daily constraints on the number of queries) for each \((\text{word}, \text{context})\) pair to determine the number of occurrences involving each pattern containing the \((\text{word}, \text{context})\). I define the pattern as the constructed relationship without the word or context; i.e., “is” or “has a.” For example, for the \((\text{context}, \text{word})\) pair of \((\text{email}, \text{person})\), the constructed relationship with the “has” pattern would be “person has email.” The queries are always enclosed in quotes, so that the search engine finds the full uninterrupted phrase on each page; otherwise Yahoo (and Google) can also match individual words of the search phrase within the same page, instead of just the whole phrases.

The patterns are chosen so that the resulting pattern results in a fundamental semantic relationship between the word and concept to minimize false positives. Also, the patterns must occur freely in natural language so that high occurrence frequency in pattern correlates with concept, as the method inherently assumes. After experimentation on the training data, patterns used for the results in this paper include “’s” (possessive case), “has (a),” “in (a),” “with (a),” and “of (a)” patterns.

One issue arising from this framework is how to determine what constitutes a frequent enough occurrence of the pattern. For example, are three occurrences of the “person’s tourists” pattern found online sufficient to conclude that the two words are semantically related such that they are in the same context, or is over 600 occurrences for “person has an office” not yet sufficient? This issue is addressed in the algorithm by
using the mutual information of the context in question with the corresponding pattern, instead of using raw occurrence numbers. The method finds the mutual information of the context concatenated with pattern concatenated with word. Any mutual information score over one implies that there are more patterns found in the text than expected by the individual general frequencies and it can be concluded that the presence of the pattern influences the word in question.

Another issue from this framework is choice of the specific patterns to be used. One seemingly feasible and elegant idea is to use patterns that underlie fundamental semantic relationships. Examples would be hypernym-hyponym relationships (a superclass of – a subclass of, respectively), holonym-meronym relationships (a whole of – a part of, respectively), synonym relationships, and so on. A problematic issue with these is that is and is-a part-of relationships is can give artificially high or low frequencies, resulting in high false positive or false negative percentages for context determination. For example, searching for “war in the context of religion” (with the aim of retrieving all mentions of war motivated or in association with religion) and using the is a pattern may determine a sentence that contains the word science to be contextually related to religion because the pattern science is religion occurs with very high frequency, if not just in a metaphorical sense.

There are some patterns that convey necessary, fundamental relationships but need to be expressed in popular, conversational ways that would lead to high frequencies. For example, “person has email” and “person’s email” both convey the has-a relation, but choosing the latter gives many more occurrences within written text on the internet.
4.3 Algorithm

Given a context and query, the method gets the first 50 web page results from the Yahoo search engine on the query without the context, using the Yahoo API. The method then extracts each word from each sentence from each webpage and applies the fixed patterns involving the context and that word. The algorithm then finds the mutual information of the [context] and the [pattern joined with the word]. The algorithm returns a positive context determination if there is a sentence with a sufficiently high mutual information score in the page. The algorithm pseudocode is shown in Figure 4.1. Determining such a threshold for the mutual information school was determined on training examples described in more detail in the Experiment Section of Chapter 5.

```plaintext
CONTEXTUAL_SEARCH (query, context)
{
    query Yahoo Search API on query and get top 50 web pages
    For each web page:
        For each sentence in web page:  // sentence ends in period
            For each non-stopword of sentence:
                For each pattern:
                    Form relationship = context + pattern + word
                    query Bing Search API for number of hits containing the pattern in full
                    calculate mutual information (MI score) of pattern
                    Calculate final (joint) MI score of all patterns for the word
                    IF (the final MI score > 1.0) AND
                        NOT(current word is a non-stopword with a final MI score < 1.0)
                        PRINT result

}
```

Figure 4-1. Pseudocode for contextual search
CHAPTER 5
EXPERIMENT AND RESULTS

5.1 Methodology

A human study was used to compare the results of the contextual search algorithm to human judgment. An alternative means of gauging the efficacy of the algorithm is to train and test the algorithm on data accepted by an established community of researchers in contextual search. Unfortunately, there exists no such community of contextual search researchers; there are many data sets and conferences devoted to finding the relevancy of search results to general search queries (Voorhees 2005, Peters 2009, Cando 2008, Majumder 2008), but there is currently no such organized compendium where the search queries designates a specific context. Moreover, the most similar research already done was a contextual search engine by Yahoo that was also validated using a human subject study (Kraft 2005).

The human study consisted of 37 participants, with a total of 11 queries. Each query consisted of a search for a certain term within the context of another. An example of such a query is “Beetle in the context of car.” The set of instructions given for an example query is given in Figure 5-1. The queries were divided into groups of two or three among all the participants, in order to promote successful completion among the participants. The specific queries that each participant received was determined randomly among the eight test queries. The participants were students gathered across three computer science classes, composing the full range of beginning, intermediate, and advanced courses. These courses were during the University of Florida Summer 2009 term: CGS2531 Problem Solving Using Computer Software, CIS3023 Introduction to Java Programming, and COP3530 Data Structures and Algorithm Analysis. The participants were all volunteers offered course extra credit and under the guidelines and
approval of the University of Florida Institutional Review Board. Each participant was given their list of queries, along with web pages containing the key word in the query, but not necessarily in the given context, and ask to rate the web page's relevance to the query using a Likert scale from 1 to 5 for relevancy. The eleven queries used were:

1. "mouse in the context of novel"
2. "Marlon Brando in the context of father"
3. "Vinh Nghiem in the context of person"
4. "war in the context of religion"
5. "flash superhero in the context of cartoon"
6. "Magic Johnson in the context of player" (not coach)
7. "scandal in the context of politics"
8. "love in the context of a song"
9. "Beetle in the context of a car"
10. "Michael Phelps in the context of a student"
11. "death in the context of disease"

The eleven queries were chosen with a varying range of difficulty for the conceptual search problem. Queries that are very difficult contain contexts that do not ordinarily appear as a string match on the page. Such examples are “Vinh Nghiem in the context of person,” “Marlon Brando in the context of a father,” “love in the context of a song,” and “Michael Phelps in the context of a student.” Michael Phelps, for example, normally is not viewed in the context of a person, but rather as a swimmer, and so pages that include the word “student” with “Michael Phelps” do not represent Phelps himself in that context, but rather have him addressing other students motivationally, for example. These examples are difficult because the context word cannot be found by
the contextual search algorithm through a direct string match. Rather, other words belonging in the context must be discovered. Queries with more intermediate difficulty include contexts that both may and may not appear as direct string matches. Examples are “war in the context of religion,” “death in the context of disease,” and “Magic Johnson in the context of basketball player (not coach).” To find war in the context of religion, for example, both “war” and “religion” may appear on the page, or “war” may appear with “crusades,” for example. Queries with low level of difficulty for contextual search have contexts that almost invariably appear with the query. Such examples are “scandal in the context of politics” since both words will almost always appear on the same page for contextually relevant pages.

The responses for each query and each participant were converted mathematically into high-dimensional vectors, with each vector containing the participant’s Likert relevancy score for each web page for that query. A randomized block ANOVA was performed on all sets of vectors to eliminate participant bias and ensure inter-rater reliability. That is, for each query, participants were removed who gave randomized or indeliberate answers reflected with a high F-score in the ANOVA analysis. For each query, a single vector was then formed comprised by the means of all vectors of that query. Each component of each vector was then converted to a binary scale to match the binary scale used by the program. The program was run on all 11 queries and was trained on the first three queries only. The parameters, namely the patterns to be used as well as the mutual information threshold, were chosen based on these training queries. The program was then tested on the remaining eight queries, whereby no changes in the parameters were done. The correlation between the program results and participant results was then found using a Spearman Rank
Correlation Coefficient. The precision and recall of the contextual search algorithm was computed by using the human judgments for each web site of each query as the ground truth. Since the Yahoo search engine API is used, and the first 50 web page results are given for every query, the document domain is tractable. These precision and recall results were compared to those of the Google search engine, again using the human judgment results as ground truth.

The purpose of this page of links is for you to determine whether each link contains any information that is relevant to the query "war in the context of religion". That is, we would like to determine the web pages that mention a "war" in the context of religion; i.e., wars fought for religious reasons, holy wars, the crusades, etc.

Please mark the relevancy of each link to the query on the survey page that contains the same query and links.

For each web page, please give a score of:

  1 for Strongly Unrelevant
  2 for Unrelevant
  3 for Neither relevant nor irrelevant
  4 for Relevant
  5 for Strongly relevant

In rating relevancy of a keyword to a given context; i.e., "war" to "religion", it may be extremely helpful to use the find function on your browser (Ctrl F) on each web page to highlight all instances of the keyword ("war"). You can then quickly scan the neighboring areas around the highlighted keyword to determine if it fits the context ("religion").

Please determine if the following web pages have any information relevant to the above query:

1. ______ http://en.wikipedia.org/wiki/War
2. ______ http://en.wikipedia.org/wiki/Korean_War
3. ______ http://plato.stanford.edu/entries/war/
4. ______ http://www.answers.com/topic/war
5. ______ http://www.netflix.com/Movie/War/70061480
6. ______ http://www.wartheband.com/
7. ______ http://en.wikiquote.org/wiki/War
8. ______ http://www.addictinggames.com/territorywar.html
10. ______ http://www.warthefilm.com/
11. ______ http://dictionary.reference.com/browse/war?jss=0

Figure 5-1. Directions for sample query
5.2 Data Analysis

The method was trained on three sample queries; namely, (1) “Vinh Nghiem in the context of person,” (2) “war in the context of religion,” and (3) “mouse in the context of novel.” The parameters, namely, the mutual information threshold as well as determination of reliable patterns, were chosen and tested on the eight remaining training samples.

The F-score values for the ANOVA are shown in Figure 5.1. All values are within a 0.05% chance of rejecting the null hypothesis that the included participants within
each query have significant variation. This suggests inter-rater reliability and that participants that answered randomly or carelessly were ignored.

The Spearman Rank coefficients are shown in Figure 5.2. These measures reflect the correlation that the human judgment of the mean participant for each query has with the contextual search algorithm result for that query. All eleven queries tested have correlations near one, with the lowest being 0.80 and with correlations of most queries above 0.9. This gives strong evidence that the contextual search algorithm performs in line with human perception.

For each query, there is an associated Table (Table 5-1 through Table 5-11) to show the precision and recall data comparing that of Google to the currently proposed approach. The output from the algorithm for each query is shown in Figure A-2 through Figure A-12. In addition, Figure 5-3 is a bar chart showing the average precision and recall data for all three queries comparing Google to the proposed method. The difference in these quality measures between Google and the proposed method are statistically significant.

Some notes follow to explain the collection of the recall and precision measurements. The Google Top 50, the only other sentences that exist on the internet are the ones found with the proposed approach. Of course, there must be at least this many additional sentences on the entire internet. The proposed approach was taken from the first 50 results returned by the Yahoo Search API.

Table 5-1. Comparison of results for the query “Vinh Nghiem in the context of person” (Vinh Nghiem mentioned as a person and not a pagoda, for example).

<table>
<thead>
<tr>
<th></th>
<th>Google</th>
<th>Proposed Approach</th>
</tr>
</thead>
<tbody>
<tr>
<td>Precision</td>
<td>0%</td>
<td>100%</td>
</tr>
<tr>
<td>Recall</td>
<td>0%</td>
<td>100%</td>
</tr>
</tbody>
</table>
Table 5-2. Comparison of results for the query “war in the context of religion” (mentions of war relating to religion).

<table>
<thead>
<tr>
<th></th>
<th>Google</th>
<th>Proposed Approach</th>
</tr>
</thead>
<tbody>
<tr>
<td>Precision</td>
<td>80%</td>
<td>92.9%</td>
</tr>
<tr>
<td>Recall</td>
<td>50%</td>
<td>100%</td>
</tr>
</tbody>
</table>
Table 5-3. Comparison of results for the query “mouse in the context of novel” (mentions of a mouse as a character in a novel)

<table>
<thead>
<tr>
<th></th>
<th>Google</th>
<th>Proposed Approach</th>
</tr>
</thead>
<tbody>
<tr>
<td>Precision</td>
<td>10%</td>
<td>100%</td>
</tr>
<tr>
<td>Recall</td>
<td>66.6%</td>
<td>100%</td>
</tr>
</tbody>
</table>

Table 5-4. Comparison of results for the query “Marlon Brando in the context of father”

<table>
<thead>
<tr>
<th></th>
<th>Google</th>
<th>Proposed Approach</th>
</tr>
</thead>
<tbody>
<tr>
<td>Precision</td>
<td>10%</td>
<td>88.9%</td>
</tr>
<tr>
<td>Recall</td>
<td>40%</td>
<td>100%</td>
</tr>
</tbody>
</table>

Table 5-5. Comparison of results for the query “Flash in the context of superhero and cartoon”

<table>
<thead>
<tr>
<th></th>
<th>Google</th>
<th>Proposed Approach</th>
</tr>
</thead>
<tbody>
<tr>
<td>Precision</td>
<td>8%</td>
<td>100%</td>
</tr>
<tr>
<td>Recall</td>
<td>31%</td>
<td>100%</td>
</tr>
</tbody>
</table>

Table 5-6. Comparison of results for the query “Magic Johnson in the context of basketball player (not coach)”

<table>
<thead>
<tr>
<th></th>
<th>Google</th>
<th>Proposed Approach</th>
</tr>
</thead>
<tbody>
<tr>
<td>Precision</td>
<td>96%</td>
<td>100%</td>
</tr>
<tr>
<td>Recall</td>
<td>66.6%</td>
<td>100%</td>
</tr>
</tbody>
</table>

Table 5-7. Comparison of results for the query “scandal in the context of politics”

<table>
<thead>
<tr>
<th></th>
<th>Google</th>
<th>Proposed Approach</th>
</tr>
</thead>
<tbody>
<tr>
<td>Precision</td>
<td>90%</td>
<td>100%</td>
</tr>
<tr>
<td>Recall</td>
<td>100%</td>
<td>100%</td>
</tr>
</tbody>
</table>

Table 5-8. Comparison of results for the query “love in the context of a song”

<table>
<thead>
<tr>
<th></th>
<th>Google</th>
<th>Proposed Approach</th>
</tr>
</thead>
<tbody>
<tr>
<td>Precision</td>
<td>100%</td>
<td>100%</td>
</tr>
<tr>
<td>Recall</td>
<td>75.8%</td>
<td>100%</td>
</tr>
</tbody>
</table>

Table 5-9. Comparison of results for the query “Beetle in the context of car”

<table>
<thead>
<tr>
<th></th>
<th>Google</th>
<th>Proposed Approach</th>
</tr>
</thead>
<tbody>
<tr>
<td>Precision</td>
<td>100%</td>
<td>100%</td>
</tr>
<tr>
<td>Recall</td>
<td>100%</td>
<td>100%</td>
</tr>
</tbody>
</table>

Table 5-10. Comparison of results matching the query “Michael Phelps in the context of student”

<table>
<thead>
<tr>
<th></th>
<th>Google</th>
<th>Proposed Approach</th>
</tr>
</thead>
<tbody>
<tr>
<td>Precision</td>
<td>2%</td>
<td>75%</td>
</tr>
<tr>
<td>Recall</td>
<td>16.6%</td>
<td>100%</td>
</tr>
</tbody>
</table>
Table 5-11. Comparison of results matching the query “death in the context of disease”

<table>
<thead>
<tr>
<th></th>
<th>Google</th>
<th>Proposed Approach</th>
</tr>
</thead>
<tbody>
<tr>
<td>Precision</td>
<td>100%</td>
<td>85.7%</td>
</tr>
<tr>
<td>Recall</td>
<td>100%</td>
<td>100%</td>
</tr>
</tbody>
</table>

Table 5-12. Final F Score values for all participants within each query (all values are within a 0.05% chance of rejecting the null hypothesis that the included participants within each query have significant variation)

<table>
<thead>
<tr>
<th></th>
<th>F Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.851963</td>
</tr>
<tr>
<td>2</td>
<td>0.42736</td>
</tr>
<tr>
<td>3</td>
<td>0.34457</td>
</tr>
<tr>
<td>4</td>
<td>0.69317</td>
</tr>
<tr>
<td>5</td>
<td>0.72379</td>
</tr>
<tr>
<td>6</td>
<td>0.386932</td>
</tr>
<tr>
<td>7</td>
<td>0.040699</td>
</tr>
<tr>
<td>8</td>
<td>0.23748</td>
</tr>
<tr>
<td>9</td>
<td>0.760048</td>
</tr>
<tr>
<td>10</td>
<td>0.33673</td>
</tr>
<tr>
<td>11</td>
<td>0.099808</td>
</tr>
</tbody>
</table>
Figure 5-2. Spearman Correlation Coefficient of the correlation of the contextual search engine result to the mean participant score for each query. A Spearman coefficient closer to +1 and -1 represent positive and negative correlations, to human judgment, respectively, while 0 represents no correlation.

Figure 5-3. Comparison of average results for contextual search queries.
5.3 Discussion of Results

In the ANOVA analysis, for each query, a maximum of two participants ended up being ignored for each query to ensure that the F-scores resulted in a maximum 0.0005 confidence in acceptance of the null hypothesis that all participants within a given query had an unacceptable variance within the group. This reflected the assurance that participants that gave random or careless responses were ignored.

The Spearman Rank coefficients reflect the correlation that the human judgment of the mean participant for each query has with the contextual search algorithm result for that query. All eleven queries tested have correlations near one, with the lowest being 0.80 and with correlations of most queries above 0.9. This gives strong evidence that the contextual search algorithm performs in line with human perception.

The data shows precision and recall results for different contextual queries. The proposed approach was taken from the first 50 results returned by the Yahoo Search API. Because of this tractable set of documents, precision and recall were determined conclusively for the data.

The Google Results are determined from only the top 50 results returned, since the quality of the search return diminishes considerably afterwards. For the Google recall measurement, it is impossible to determine how many conceptually correct sentences exist on the internet, so it is assumed that besides the ones returned by the Google Top 50, the only other correctly identified sentences that exist on the internet are the ones found with the proposed approach. Of course, there must be at least this many additional sentences on the entire internet. Thus, the data for the Google recall represents an upper bound. In reality, it is expected the Google recall is much lower since two-thirds of the results correctly identified by the proposed method did not
contain any of the words in the query—results for which Google surely would not have found.

The method works by using pre-constructed patterns. For every (word, context) pair, the method queries the Yahoo Search Engine API to determine the number of occurrences involving each pattern containing the (word, context). The pattern is the constructed relationship without the word or context; i.e., “is” or “has a.” For example for (email, person), the constructed relationship with the “has” pattern would be “person has email.” The method then finds the mutual information of the context+pattern skeleton and word. It is important to note that the queries are always enclosed in quotes, so that Yahoo finds the phrase in full on each page; otherwise, Yahoo (and Google) only match individual words of the search phrase.

A large number of patterns have been experimented with including “’s” (possessive case), “has,” “has a,” “with,” “with a,” “is,” “is a,” “is defined by,” “such as,” “is a type of,” “is a part of,” “for example,” “is an example of,” a null pattern (i.e., in “height person” there is a null pattern skeleton joining “height” and “person”). The successful pattern skeleton will express a fundamental semantic relationship between the word and context, as well as be expressed in language naturally enough to frequently occur.

A pattern found that matches both aims is the “’s” (possessive case). The pattern: context’s word, such as “person’s office,” effectively communicates ownership of the word by the context, while being a casual, popular means of expressing the relationship.

The “is a” relationship, while it can express a fundamental semantic relationship, is also often used as a figure of speech, to speak in metaphor (i.e., “science is a religion”). As a consequence, the “is a” relationship gives a very large number of hits for
most queries, leading to a high mutual information, and ultimately leading to high false positives.

Likewise, the null pattern to detect, for example, whether one word is a fundamental adjective of another (i.e., in “sacred religion”) gives a high number of hits because punctuation marks such as the colon and period are ignored by the search engine on the page. This results in unintended hits and an artificially high count of the pattern, again leading to high false positives.

Conversely, there are patterns that express fundamental semantic relationships but in themselves are not as naturally spoken and give artificially low hits. Examples are “with a” and “has a,” such that “a person’s email” is much more naturally occurring than “a person has email.” Consequently, these patterns are used the method in addition to patterns that occur more naturally in text.

Even though the existence of some words in a sentence will almost guarantee context (i.e., “holy” for the context of religion), one reality of language is that a word could have a fundamental relationship to not exclusively one word, but to others as well, such that its existence in the sentence does not constitute the sentence’s membership in a context. For example, “theories” could be about many contexts, whether it be science, religion, politics, and so on. This motivates the second part of the method, which filters the possibility of another context. If the phrase “scientific theories” is encountered while the desired query is war in the context of religion, and if, for example, the pattern is “s’, then the mutual information score of “religions’s theories” is found relative to the baseline score of “religion’s war.” Without this addition to the method, there were about 15% more false positives found.
A subset of the issue just mentioned is that some words common to more than one context are so unspecific they are considered unimportant. They are sometimes coined stopwords in search engine terminology since they are ignored by search engines. Examples of such words are “the,” “to,” “all,” etc. If left in the sentence, they will result in artificially high hits. One approach around the issue is to keep a list of stopwords, which are widely available. A second approach is to compare the mutual information score of the desired context to the baseline score of the original query by taking the ratio of the two quantities. This method uses both methods to filter out only context-specific words.

Another limitation in the method is that it filters from the first 50 Yahoo search engine results, rather than the whole internet. This allows tractability in measuring recall results and also is currently sufficient to assess the quality of the method. For practical, real-time systems, however, the method can be used by a web crawler and indexer to actually go through all pages that contain a certain query term that needs to be in a certain context, instead of the first 50 from a given search engine.
CHAPTER 6
ISSUES FOR IMPROVEMENT AND FUTURE DIRECTIONS

6.1 Issues for Improvement

While the method correctly determines context for most sentences in natural language, it is unable to examine the structure of the sentence. For example, finding the query “Ann” in the context of “mother” will yield a positive context results for both sentences “Ann drove her daughter to school” and “Ann drove her sister’s daughter to school,” where only the former should be true. Further work needs to be done to extract structure in the existing framework.

Another area for improvement in the algorithm is to perform named-entity extraction as a preliminary step before applying the current algorithm. Named-entity extraction identifies a contiguous sequence of words that represent a single entity. For example, if the user is searching for a context of “location” and one of the candidate words on a web page is “New York,” then the algorithm would determine whether the entity “New York” belongs in the “location” context, rather than just the words “New” and “York” individually.

Currently, the algorithm determines whether words in the same sentence as a desired search term belong in a given context. Furthermore, although the algorithm measures similarity in fundamental semantic relationships, it does not examine sentence structure. For example, for the search “Marlon Brando in the context of father,” if the sentence “Marlon Brando’s sister had three children” were encountered, the algorithm would make a positive context determination on the sentence because “children” has a high final mutual information score with “father.” A future improvement to the algorithm would be to also incorporate the processing of sentence structure into the mutual information scores.
Currently, both precision and recall rates of over 94% and correlations of over 0.8 to human judgment suggest promising performance. A preliminary named-entity extraction step, as well as an attempt at sentence structure analysis would only improve performance further.

### 6.2 Future Directions

In this research, contextual search determination was done on the first 50 web pages that matched a desired search term. This domain restriction was done to maximize participant cooperation in judging the relevance of the pages and to make the computation measures of precision and recall tractable. Because the method’s context determination approach relies on a final mutual information score, that in turn, relies only on a page count with hits for a specific word, the method can scale immediately onto the entire internet. Because of this lack of restriction in domain and the method’s identification of fundamental semantic relationships between words, it can be utilized to find new relationships and knowledge in ontology generation. Furthermore, the algorithm can be applied directly to existing web crawlers in implementing practical, real-time contextual search systems.

The method can be used for ontology generation to discover and store new repositories of fundamental relationships between words. Just by doing a routine search engine crawl of web sites, relationships with high mutual information scores can automatically be recorded into a repository. For example, “a person has an office” and “a novel is a book” would be such relationships recorded. These relationships could be valuable for data mining and information extraction tasks.

Currently, given the context, the method gets the first 50 results from the Yahoo search engine on the query without the context. It then extracts a word from every
sentence from every webpage. Next, it applies the fixed patterns involving the context and that word and finds the mutual information of the \[context\] and the \[pattern joined with the word\]. Finally, it returns a positive context determination if there is a sentence with a high final mutual information score in the page. Obviously, this amount of computation would be unacceptable in a real-time application. A feasible solution could be that the first set number of runs would be considered the training phase, so that every time a word is encountered that has a very high MI score and thus is in context with the given context, the web page is stored in a \[(page, sentence, context)\] index to denote that somewhere in the page, a context exists through a certain sentence. Then, in real-time, the desired web pages represent the intersection of the index of pages that match the context and the index of pages that match the search term. In this way, though, popular contexts need to be anticipated in the training process. One possibility of making the guesswork easier is to use the most popular categories from the Open Directory Project.

The research could shed insight and new research into the inherent nature of human context determination. The extent and specific role of mutual information and lexico-syntatic patterns in linguistics and natural language processing can be explored.
I proposed a novel approach for contextual search, using mutual information and lexico-syntactic patterns. I implemented the algorithm and tested it on examples that are representative of the gamut of contextual search difficulty. I have also performed a human subject study to test the efficacy of this method. Results on contextual search queries with our subjects have been highly promising, with precision and recall rates both over 94%, compared to Google’s precision and recall rates of under 55%. The human subject experiment with sample queries shows consistently strong correlation (0.8-1.0) of the contextual search algorithm to human judgment of context. This research allows for contextual search on the internet, which is more accurate than existing means. It would also lead to insight on the effect the mutual information of words with lexico-syntactic patterns has on the determination of context in natural language processing.

There are several future directions this research can take. First, a practical, mainstream application can be developed, which indexes a list of pages belonging to a given context, and in-real-time, can instantly return a list of pages that belong in a given context. Secondly, the method can be used for ontology generation to discover and store new repositories of fundamental relationships between words. Lastly, the research could shed insight and new research into the inherent nature of mutual information with lexico-semantic patterns for context determination in linguistics and natural language processing.
**APPENDIX**

Figure A-1. Precision and recall values for different methods of calculating composite mutual information score across multiple patterns.

```plaintext
Output for sentences within web pages answering the query "Vinh Nghiem in the context of person":
http://www.cise.ufl.edu/~cgs2531/faq.php

MWF 5 CSE E222 3592 Vinh Nghiem CSE 502 392-8968 vinh.cgs@gmail.com
Where are the instructors' offices and what are the instructors' hours? Name Office Phone Office Hours Rory De Simone CSE 512 Email MWF 5 Shibdas Bandyopadhyay CSE 502 392-8968 MWF 8 Udayan Kumar CSE 502 392-8968 MWF 7 Sungwook Moon CSE 502 392-8968 MWF 8 & 9, F7 Vinh Nghiem CSE 502 392-8968 MWF 9 Vishak Sivakumar CSE 502 392-8968 T 4 Th 3, 4
What are the exam dates? All exam dates are listed on the course calendar.
*** Word = office, Context = Person, Final MI score = 1.9275572650409654
```

Figure A-2. Output for results matching the query “Vinh Nghiem in the context of person”

```plaintext
Output for sentences within web pages answering the query “War in the context of Religion”:
0 http://en.wikipedia.org/wiki/War

Fornari thought that war and violence develop out of our "love need": our wish to preserve and defend the sacred object to which we are attached, namely our early mother and our fusion with her.
*** Word = sacred, Context = Religion, Final MI Score = 102.84749405173076
For such a war, even the biblical exemptions (see above) had to be suspended: a newly married man must take part in the fighting as a sacred duty; and Sabbath prohibitions could also be waived in the interests of national defense (Tosef).
*** Word = sacred, Context = Religion, Final MI Score = 102.84749405173076
It sees wars as imperial ventures to enhance the power of the ruling class and divide the proletariat of the world by pitting them against each other for contrived ideals such as nationalism or religion.
*** Word = religion matches context
```

Figure A-3. Output for results matching the query “war in the context of religion”
Throughout history war has been the source of serious moral questions. Although many ancient nations and some modern ones have viewed war as noble, over the sweep of history, concerns about the morality of war have gradually increased. Today, war is seen by some as undesirable and morally problematic.

Aristotle What difference does it make to the dead, the orphans and the homeless, whether the mad destruction is wrought under the name of totalitarianism or the holy name of liberty or democracy? Mohandas Gandhi What is War Crimes when war in itself is a crime? Calvin Austin When people speak to you about a preventive war, you tell them to go and fight it.

The wars and battles of the Bible had faded into a distant past, and now the struggle against Rome was also becoming a vague memory. While this concept of a "holy war" proved vital in Hasmonae times (cf.3:21), most of the sages, when the Jews no longer enjoyed independence, had no firsthand knowledge of warfare.

sees wars as imperial ventures to enhance the power of the ruling class and divide the proletariat of the world by pitting them against each other for contrived ideals such as nationalism or religion.

However, the 1982 Lebanon War produced a division of religious opinion, some authorities defining it as a mili?emet ha-reshut, an optional war, undertaken for purposes beyond legitimate defense.

This finds religious expression in the directives issued by the military rabbinate governing conduct in time of war and the waiving of Sabbath prohibitions, etc.,

Type Example Extortionate Pecheneg and Cuman forays on Rus in 9th-13th centuries Aggressive the wars of Alexander the Great in 326-323 BC Colonial Sino-French War Imperial Rebellion Algerian War Religious Crusades Dynastic War of the Spanish Succession Trade Opium Wars Revolutionary French Revolutionary Wars Guerrilla Peninsular War Civil Spanish Civil War Secessionist American Civil War Nuclear There is only one recorded use of nuclear weapons; during World War II, the U.S. Capulla, War Wolf, University of Pili (2008) See also War portal Peace General reference articles Feud Undeclared war Colonial war Religious war Breakaway states

Figure A-3. Continued
Holy war.

10 http://www.imdb.com/title/tt0372873/
The very idea of a selfish pacifist simply does not ring true: many pacifists have, historically, paid a very high price for their pacifism during wartime (through severe ostracism and even jail time) and their pacifism seems less rooted in regard for inner moral purity than it is in regard for constructing a less violent and more humane world order.

Figure A-3. Continued

OUTPUT FOR SENTENCES WITHIN WEB PAGES ANSWERING THE QUERY “MOUSE IN THE CONTEXT OF NOVEL”:

http://www.thefreedictionary.com/mouse
The Yellow Fairy Book by Lang, Andrew View in context Some said this, and some said that; but at last a young mouse got up and said he had a proposal to make, which he thought would meet the case.

http://encyclopedia2.thefreedictionary.com/mouse
The Yellow Fairy Book by Lang, Andrew View in context Some said this, and some said that; but at last a young mouse got up and said he had a proposal to make, which he thought would meet the case.

Figure A-4. Output for results matching the query “mouse in the context of novel”

OUTPUT FOR SENTENCES WITHIN WEB PAGES ANSWERING THE QUERY “MARLON BRANDO”:

http://www.imdb.com/name/nm0000008/bio
Acting was a skill he honed as a child, the lonely son of alcoholic parents

Eldest son Gary Brown was arrested for murdering his half-sister's boyfriend Dag Drollet in 1990. He was sentenced to 10 years in prison in March 1991 and released in January 1996.

Youngest of three children.

His son Miko C. Brando was once a bodyguard for Michael Jackson.

Cheyenne's son Tuki Brando models for Versace menswear

Figure A-5. Output for results matching the query “Marlon Brando in the context of father”
His eccentric lifestyle, though, kept him in the press: his (at least) nine children by various wives and companions; his 1972 Oscar refusal delivered by faux Native American Sasheen Littlefeather, in protest of Hollywood's depiction of the indigenous tribes; offbeat and outrageous on-set behavior; the killing of daughter Cheyenne's fiance by son Christian; Cheyenne's subsequent suicide at age 25; and an amusing televised incident where he playfully insisted interviewer Larry King kiss him on the lips.

http://www.respectance.com/marlon_brando
Brando's mother was a kind and talented woman, although she suffered from alcoholism and was frequently absent in Brando's childhood.

http://www.tabloidcolumn.com/marlon-brando-died.html
Marlon Brando cut out of his will one of his teenage grandsons and an adopted daughter, according to a copy of the document released yesterday.

http://www.wordiq.com/definition/Marlon_Brando
The number of children he had is still in dispute, although he recognized 11 children in his will;

Figure A-5. Continued

OUTPUT FOR SENTENCES WITHIN WEB PAGES ANSWERING THE QUERY “FLASH IN THE CONTEXT OF SUPERHERO AND CARTOON”:

http://www.kidzworld.com/article/8032-top-10-cartoon-superheroes
We're taking a look at our top 10 cartoon superheroes of all time.

http://en.wikipedia.org/wiki/Flash_(comics)
The Wally West version of the Flash (but with many elements of Barry Allen's story) is featured in the animated series Justice League.

http://www.flashsuperhero.com/
ClassicWBTV The Flash TV Show | The Flash Episodes | Veoh Sign Up Log In English Deutsch Français 日本語 Add Video My Veoh Videos TV Shows Movies Music Channels Groups Featured Check out the new Bookmarking Tool and Playlist features! LARGE SMALL Video Size: The latest version of Adobe Flash Player is required to watch video.

Figure A-6. Output for results matching the query "Flash in the context of superhero and cartoon"
In his final high school season, Johnson led Lansing Everett to a 27-1 win-loss record while averaging 28.8 points and 16.8 rebounds per game, and took his team to an overtime victory in the state championship game.

He also became one of four players to win NCAA and NBA championships in consecutive years.

Basketball Hall-of-Famer, first rookie ever to win the Finals MVP Award (1980)

Earvin “Magic” Johnson Few athletes are truly unique, changing the way their sport is played with their singular skills.

Earvin Johnson Jr. (Magic) Inducted into Hall of Fame as Player in 2002.

Earvin Johnson Jr., better known as Magic Johnson, is a Hall of Fame basketball player and an executive for the Los Angeles Lakers.

Three times Magic was the MVP of the league and three times the Finals MVP.
Johnson played in 12 All-Star Games, was the league's Most Valuable Player three times (1987, '89 and '90) and was a member of the USA's "Dream Team," gold medalists in the 1992 Olympics (after he'd announced his retirement).

Johnson's team won convincingly, and he was named Most Valuable Player in the N.C.A.A.

When the shooting stopped, Magic had claimed three Most Valuable Player Awards and five championship rings, while "Larry Legend" had three MVP Awards and three championships.

Some scandals are broken by whistleblowers who reveal wrongdoing within organizations or groups, such as Deep Throat (William Mark Felt) during the 1970s Watergate scandal.

Western world In the United States and Canada, scandals, particularly political ones, are often referred to by adding the suffix "-gate" to a word connected with the events, recalling the Watergate scandal, such as "Nannygate".

A publicized incident that brings about disgrace or offends the moral sensibilities of society: a drug scandal that forced the mayor's resignation.

The public no longer expects movie stars, politicians, athletes, chief executive officers, or even the president of the United States to remain free of scandal.

The Iran-Contra scandal of the mid-1980s not only destroyed the careers of several high-ranking officials in the Ronald Reagan administration, it caused a national crisis of confidence in the democratic system.
A number of scandals during the presidency of Bill Clinton (1993–2001), from the Whitewater real estate scheme to the president's affair with White House intern Monica Lewinsky, revealed the way public opinion about misconduct had changed.

Some scandals are broken by whistleblowers who reveal wrongdoing within organizations or groups, such as Deep Throat (William Mark Felt) during the 1970s Watergate scandal.

Western world In the United States and Canada, scandals, particularly political ones, are often referred to by adding the suffix "-gate" to a word connected with the events, recalling the Watergate scandal, such as "Nannygate".

Read more Related answers What was the ENRON Scandal? Cause of the watergate scandal? What was the blacksox scandal? What was the cause of the Watergate scandal? » More

Figure A-8. Continued

Figure A-9. Output for results matching the query “Love in the context of a song”

http://www.answers.com/topic/love
Because of the complex and abstract nature of love, discourse on love is commonly reduced to a thought-terminating cliché, and there are a number of common proverbs regarding love, from Virgil's "Love conquers all" to the Beatles' "All you need is love."

Jimi Hendrix The woman was made of a rib out of the side of Adam; not made out of his head to rule over him, nor out of his feet to be trampled upon by him, but out of his side to be equal with him, under his arm to be protected, and near his heart to be beloved.

John Lennon Love means having to say You're sorry every five Minutes John Lennon Everything is clearer when you're in love.

John Lennon How can I give love when I don't know what it is I'm giving? John Lennon, 'How?' Love is the flower you've got to let grow.

John Lennon Love is a promise, love is a souvenir, once given never forgotten, never let it disappear.

John Lennon All you need is love.
John Lennon/Paul McCartney And in the end the love you take is equal to the love you make.

*** Word = Mccartney, Context = song, Final MI score = 16.78805902668864
*** Word = Lennon, Context = song, Final MI score = 109.18052681132959

U2 True love never can be rent, but only true love can keep beauty innocent.
*** Word = song, Context = song, Final MI score = 46.267370948881045

U2 You say love is a temple, love a higher law.
*** Word = U2, Context = song, Final MI score = 46.267370948881045

U2 [edit] V The more I think about it, the more I realize there is nothing more artistic than to love others.
*** Word = U2, Context = song, Final MI score = 46.267370948881045

Figure A-9. Continued

http://www.answers.com/topic/beetle
Beetles have hemolymph instead of blood, and the open circulatory system of the beetle is driven by a tube-like heart attached to the top inside of the thorax.
*** Word = driven, Context = car, Final MI score = 47.800858930412886

http://www.ehow.com/how_4507300_change-oil-vw-beetle.html

Figure A-10. Output for results matching the query “Beetle in the context of car”

http://en.wikipedia.org/wiki/Michael_Phelps
In 2004, Phelps left North Baltimore Aquatic Club with Bob Bowman to train at the University of Michigan for Club Wolverine.
*** Word = university, Context = student, Final MI score = 664387.7178166816


Figure A-11. Output for results matching the query “Michael Phelps in the context of student”
Indeed a succession of lucrative sponsorship deals meant Phelps could no longer train as an amateur, so when Bowman moved to coach at the University of Michigan in 2004, Phelps followed too and trained at a local swimming club in Ann Arbor called Club Wolverine.

Between 2004 and 2008, Phelps attended the University of Michigan in Ann Arbor, Michigan, studying sports marketing and management.

Figure A-11. Continued

The leading causes of death in developed countries are atherosclerosis (heart disease and stroke), cancer, and other diseases related to obesity and aging.

But sometimes a healthy fear of death spirals out of control and becomes a phobia. With the realization that death is eminent, most people suffer from depression.

Putrefaction, bacterial and enzymatic decomposition, begins shortly after death, caused by microorganisms which enter the body.

Some African Americans expressed concern that brain death was being used to take organs prematurely from blacks for transplantation to whites.

Figure A-12. Output for results matching the query “death in the context of disease”
LIST OF REFERENCES


BIOGRAPHICAL SKETCH

Vinh Nghiem received his undergraduate degree from Harvard. He graduated with a PhD in computer engineering at the University of Florida in December 2009. His research interests lie in natural language processing, machine learning, and artificial intelligence.