THE RELATIONSHIP BETWEEN QUERY LENGTH, PARTS OF SPEECH
USAGE AND WEB SEARCH QUERY SPECIFICITY

A Thesis in
Information Sciences and Technology

by

Carolyn Theresa Hafernik

© 2013 Carolyn Theresa Hafernik

Submitted in Partial Fulfillment
of the Requirements
for the Degree of

Master of Science

August 2013
The thesis of Carolyn Theresa Hafernik was reviewed and approved* by the following:

Bernard J. Jansen
Associate Professor of Information Sciences and Technology
Thesis Advisor

Madhu Reddy
Associate Professor of Information Sciences and Technology
Graduate Program Director, Information Sciences and Technology

Gerald Santoro
Senior Lecturer, Information Sciences and Technology
Assistant Professor of Communication Arts and Sciences, College of the Liberal Arts

Michael McNeese
Professor of Information Sciences and Technology
Associate Dean for Research, Graduate Studies and Academic Affairs

*Signatures are on file in the Graduate School
ABSTRACT

Understanding the specificity of Web search queries can enable search systems to better address the underlying needs of searchers by providing them with relevant content. Identifying the specificity of Web search queries requires classifying them on a granularity spectrum from narrow to broad for a given topic. The goal of this research is to automatically classify Web search queries as either general or narrow so that the search system can treat queries within each group differently when retrieving results. I use content analysis and prior research to develop a list of nine attributes to identify the specificity of queries, and I then apply these attributes to classify 5,115 unique queries as narrow or general. I then investigate whether (1) query length and (2) part of speech can be leveraged to automatically classify queries as narrow or general. I also explore whether users show distinct patterns in their usage of the parts of speech. I find that query length and parts of speech when combined provide better identification of query specificity (80% accuracy) than either factor can alone (74% and 56% accuracy). Also, I find that I can separate users into four distinct groups based on their usage of parts of speech, indicating that modeling specificity differently for each group of users could be beneficial. Identifying the specificity of queries has implications for search engines, marketers and users.
TABLE OF CONTENTS

LIST OF FIGURES .................................................................................................................. vi
LIST OF TABLES ..................................................................................................................... vii
ACKNOWLEDGEMENTS ......................................................................................................... ix

Chapter 1  Introduction ........................................................................................................... 1
  Web Search Personalization ............................................................................................... 2
  Web Search Query Specificity .......................................................................................... 2
  Research Objectives ........................................................................................................ 4

Chapter 2  Related Work ....................................................................................................... 6
  Specificity as a Part of User Intent .................................................................................... 6
  Reformulation of Queries and Specificity .......................................................................... 8
  Ambiguous Queries and Specificity .................................................................................. 9
  Factors related to Specificity ............................................................................................ 11
    Query Length .................................................................................................................. 12
    Parts of Speech Usage .................................................................................................... 13
  Other Studies related to Specificity .................................................................................. 14
  Synthesis of Prior Work .................................................................................................... 14

Chapter 3 Methodology ......................................................................................................... 17
  Dataset .................................................................................................................................. 17
  Research Objective 1: Provide a systematic review of specificity to develop clear guidelines for identifying specificity ................................................................. 18
  Research Objective 2: Evaluate the effectiveness of query length and parts of speech in identifying the specificity of queries ......................................................... 18
  Research Objective 3: Automatically identify queries as narrow or general based on results of an analysis of query length and parts of speech for manually classified queries ......................................................... 22
  Research Objective 4: Explore whether I can automatically identify distinct user groups by their patterns of part-of-speech usage ......................................................... 23

Chapter 4  Results .................................................................................................................. 25
  Research Objective 1: Provide a systematic review of specificity to develop clear guidelines for identifying specificity ................................................................. 25
  Research Objective 2: Evaluate the effectiveness of query length and parts of speech in identifying the specificity of queries ......................................................... 28
Research Objective 3: Automatically identify queries as narrow or general based on results of an analysis of query length and parts of speech for manually classified queries ................................................................. 34
Research Objective 4: Explore whether I can automatically identify distinct user groups by their patterns of part-of-speech usage ........................................ 44

Chapter 5 Discussion ........................................................................................................ 52
   Summary of Findings ..................................................................................................... 52
   Limitations of the Research ....................................................................................... 54
   Strengths of the Research ......................................................................................... 55
   Implications ................................................................................................................ 55
   Conclusions ................................................................................................................ 57

References .......................................................................................................................... 59

Appendix A Dataset Table descriptions ........................................................................ 68
Appendix B Part-of-Speech Tags .................................................................................... 73
Appendix C Example Queries for the Most Common Parts of Speech ....................... 75
LIST OF FIGURES

Figure 4-1: Algorithm 1 using only query length ............................................................. 38

Figure 4-2: Algorithm 2 conservative approach using parts of speech with >=84% narrow ......................................................................................................................... 39

Figure 4-3: Algorithm 3 using parts of speech with >=75% narrow ......................... 40

Figure 4-4: Algorithm 4 conservative approach using both query length and parts of speech ...................................................................................................................... 41

Figure 4-5: Algorithm 5 approach using both query length and parts of speech .... 42

Figure 4-6: Patterns of usage of parts of speech for users. 1) All parts of speech. 2) Mainly Nouns. 3) Few Adverbs. 4) Mainly Nouns and Adjectives. Y-axis is the percentage of terms used by the user in the group. X-axis is groups: 1- Nouns, 2- Adjectives, 3- Adverbs, 4- Verbs, 5- Other ........................................ 47

Figure 5-1: Example possible usage of identification of web search query specificity ................................................................................................................................. 56
LIST OF TABLES

Table 3-1: The different parts of speech automatically identified and their frequency in the dataset .................................................................20

Table 4-1: Definitions and associations with specificity ..............................................26

Table 4-2: The nine attributes and example queries for each ........................................28

Table 4-3: Definitions and associations with specificity ..............................................29

Table 4-4: Effect of increasing the number of terms on the percentage of narrow and general queries classified. Chi square for number of terms * p<0.0001........30

Table 4-5: Effect of increasing number of characters on the percentage of narrow and general queries classified. Chi square for number of characters * p<0.0001, N.S. not significant ......................................................30

Table 4-6: Percentages of queries with each broad part of speech that are classified as narrow. Chi square for broad parts of speech * p<0.0001, N.S. not significant .................................................................31

Table 4-7: Percentages of queries with each part of speech that are classified as narrow. Chi square for parts of speech * p<0.0001, ** p<0.001, ***p<0.05, N.S. not significant .................................................................31

Table 4-8: Percentages of queries with each part of speech that are classified as narrow by number of terms in the query. Chi square for parts of speech * p<0.0001, **p<0.01, ***p<0.05, N.S. not significant .................................................................33

Table 4-9: Percentages of queries with each part of speech that are classified as narrow by number of terms in the query. – sample size <25. Chi square for parts of speech * p<0.0001, **p<0.01, ***p<0.05, N.S. not significant. ................. 33

Table 4-10: Statistics for part of speech usage in unique queries.................................35

Table 4-11: Summary of results of the five algorithms and their accuracy. The bolded line, Algorithm 4, gives the best result. .........................................................43

Table 4-12: Statistics for Part of Speech Usage By Users...........................................44
Table 4-13: Four clusters of users showing different patterns of parts of speech usage

Table 4-14: The average usage of the parts of speech for the four clusters of users
ACKNOWLEDGEMENTS

I would like to thank my advisor for his help and guidance, my thesis committee for thesis for their help and feedback, and my family for their support.
Chapter 1

Introduction

Today there is an unprecedented amount of information available to people via the web. Information can come in many forms: text, images, videos, or audio or a combination of different forms. With all of the information available, it is a challenge to find the specific information that meets one’s needs. Information is readily accessible, but locating information in which one is interested can be more difficult. Search engines, like Google, take queries and return results, as a list of websites that the search engine concludes contain the information the user’s query described. Still there is no guarantee that the list will contain useful results, and a user may have to go through pages of results to find the desired information. Most users are unwilling to go through pages of results as such it is important that the correct relevant results be near the top of the results list. In addition, web search engines are better at providing satisfactory results for the average user than providing results for specific individuals (Teevan, Dumais, & Horvitz, 2005a).

Moreover, there is considerable variation in what different individuals see as relevant. These realities have led to an increasing interest in personalization and into improving our understanding of users’ goals and intents. Personalization of web search has the potential to generate results that satisfy an individual’s “information need” instead of simply satisfying the need of a large group of people.
Web Search Personalization

Obtaining information about users is a central issue in information search and retrieval research today. A growing area of research is in providing targeted or personalized search results or ads. Information about users is crucial because with more user information, a search engine can provide results that are personalized for that user as opposed to results that are targeted for the typical user. Additionally, information about users aids in disambiguating queries. For instance, with information about a user’s geospatial location, the search results can be targeted for that location (e.g. a search for restaurants can return only restaurants in the user’s current location). Employing user information to target search results towards specific users is a form of personalization in which the content or appearance of a system or webpage is modified based on user characteristics such as interests, past behavior, preferences, etc. Personalization is increasingly being used as a method of improving search results (Ferragina & Gulli, 2005; Sontag, et al., 2012; Teevan, Dumais, & Horvitz, 2005b; Wedig & Madani, 2006; Xu, Wang, Zhang, & Chen, 2007).

Web Search Query Specificity

An essential component of Web search personalization is identifying the goals or tasks of searchers (Guo & Agichtein, 2010; Jansen, Booth, & Spink, 2008; Rose & Levinson, 2004). One important aspect of this is determining the user intent of Web search queries. Gonzalez-Caro et al. identify specificity as an aspect of user intent (González-Caro & Baeza-Yates, 2011; González-Caro, Calderón-Benavides, Baeza-
Web search query specificity can be defined as how narrow or general the intent and goal behind the query is. Queries have different levels of specificity. Identifying the specificity of Web search queries entails classifying them on a spectrum of narrow to broad for a given topic. Some queries may be looking for specific information while others are looking for general information. Understanding these differing levels of specificity is an important area for research because it gives pertinent information about users, their goals, and their desired results. Understanding specificity enables search engines to provide more relevant information to searchers. Therefore, detecting the specificity of Web search queries has important implications for searchers, search engines, advertisers and other content providers and is a critical area of research.

Two questions driving my research are (1) what is specificity? and (2) how can we identify it? Specificity has rarely been studied in a systematic way. In this research, we are concerned particularly with the specificity of Web search queries. The difficulties inherent in determining specificity are illustrated by in the following three queries, as an example set:

a) “restaurants”

b) “restaurants in San Francisco”

c) “Taqueria on Geary Street San Francisco.”

From the sequence of queries, one can construct a hierarchy of the queries’ specificity: (a) is least specific, (c) is most specific, and (b) is somewhere between the other two. However, constructing a hierarchy requires comparing multiple queries about similar subjects. In order to identify something as specific relatively, one needs something that is general. It is more difficult to identify the specificity of a query if one
has two queries that are unrelated to each other in terms of topic. For instance, given the queries: “book” and “California Academy of Sciences” one would probably identify the second as more specific, but one is comparing queries on unrelated topics and basing the evaluation on a subjective concept of being narrow or general.

My ultimate goal is to gain a better understanding of what query specificity means and how it can be automatically identified in order to classify queries. In this research query specificity level is defined as how narrow or general the user intent behind the query is. I develop an operationalized definition of specificity for identifying queries as narrow or general based on prior work and analysis of queries. This research uses two query characteristics, (1) parts of speech and (2) query length to automatically classify query specificity and examine specificity at the user level.

**Research Objectives**

I have four research objectives:

1) *Provide a systematic review of specificity to develop clear guidelines for identifying specificity.*

For research objective one, I examine prior work related to the specificity of Web search queries and analyze actual search queries from a transaction log to develop an operationalized definition of specificity that gives clear guidelines for its identification. I use my definition to split queries into two groups, narrow queries and general queries.

2) *Evaluate the effectiveness of query length and parts of speech in identifying the*
specificity of queries.

For research objective two, I manually label queries as either narrow or general and annotate the queries with information about two factors, their length and the parts of speech present. I then analyze what these factors can tell us about the specificity of the query. I explore whether the usage of particular parts of speech is indicative of query specificity, whether query length is indicative of query specificity and whether these factors combined can improve classification of specificity.

3) **Automatically identify queries as narrow or general based on results of an analysis of query length and parts of speech for manually classified queries.**

For research objective three, I use query length and parts of speech to develop and test five algorithms for automatically classifying Web search queries as either narrow or general. The five algorithms use different combinations of information based on part-of-speech usage, query length and a combination of the two factors.

4) **Explore whether I can automatically identify distinct user groups by their patterns of part-of-speech usage.**

For research objective four, I look at the patterns of part-of-speech usage for users and automatically place users into distinct clusters. Here, I aim to understand whether modeling specificity based on groups of users would be beneficial in order to better understand users’ specificity patterns and query intents.
Chapter 2
Related Work

Research concerning the specificity of Web search queries is a growing area that is related to several other aspects of information searching. Many factors possibly affect specificity, including query topic, the demographics of the user submitting the query, and user intent of the query. Each factor is a potential area of research on specificity. Ultimately, determining query specificity aims to reveal the underlying user goal or intent. Here, I first examine research on user intent of queries. Second, I investigate query reformulation and its relation to specificity. Third, I discuss research on ambiguous queries and how they are related to query specificity. Fourth, I review web search query research related to query length and parts of speech. Fifth, I look at other research connected to query specificity. Lastly, I integrate findings from prior research to show how they influence my research.

Specificity as a Part of User Intent

Understanding the larger concept of user intent is essential to understanding specificity. Gonzalez-Caro et al. identify specificity as one of ten aspects of user intent (González-Caro & Baeza-Yates, 2011; González-Caro, et al., 2011). Sellen et al. classified Web activities as finding, information gathering, browsing and transacting, communicating, and housekeeping (Sellen, Murphy, & Shaw, 2002). The first four of the
activities often employ search engines and the specificity of the queries entered can be related to the type of user activity. For instance, if a user is trying to complete a banking transaction, then the user has a narrow goal and thus a query related to this task would likely be narrow.

Research on user intent often focuses on three distinct user intents: navigational, informational, and transactional (Broder, 2002; Jansen, et al., 2008; Kathuria, Jansen, Hafernik, & Spink, 2010; Rose & Levinson, 2004). The intents can imply differing levels of specificity for queries. For instance, navigational or transactional queries often have narrow goals or intents and thus are narrow in specificity. Informational queries, on the other hand, can range from general to narrow. Jansen et al. (2008) propose three hierarchical levels of specificity of user intent, ranging from general at level 1 to narrow at level 3.

In addition to defining different classifications of user intent, researchers have investigated specific categories of user intent. These categories include commercial queries (Ashkan & Clarke, 2009a, 2009b; Dai, et al., 2006; Guo & Agichtein, 2010; Hu, Yang, & Li, 2008), product queries (Shen, Li, Li, & Zhou, 2009), math queries (Kamali, Apacible, & Hosseinkashi, 2012), geographic queries (R. Jones, Zhang, Rey, Jhala, & Stipp, 2008), question queries (White & Iivonen, 2001), long queries (Bendersky & Croft, 2009) and navigational queries (W. M. Lee & Sanderson, 2010).

These categories inform my research of specificity because they are often a subset of the larger concept of narrow queries. Therefore, knowledge about differences in query type has implications for specificity. For instance, question queries and navigational queries often have very specific answers that users are expecting, so are by their nature
narrow. Additionally, the different types of queries can often have even narrower categories under them. For instance, the buying funnel is often used as a way of describing the method consumers follow before purchasing items (Hotchkiss, 2004; Laycock; Nimetz, 2007). Product queries and commercial queries could potentially be classified into the stages of the buying funnel as Jansen and Schuster (2011) do for key phrases from keyword advertising campaigns, providing even more information about the specificity of the user’s goals and intents.

Reformulation of Queries and Specificity

Another area of research impacting specificity is query formulation and reformulation. Queries can be reformulated to be either more general or more narrow. Close to half of users reformulate their initial queries (Spink, Wolfram, Jansen, & Saracevic, 2001). Several studies have examined how users reformulate their queries (Jansen, Zhang, & Spink, 2007; Lau & Horvitz, 1999; Rieh & Xie, 2006). Work on query reformulation often discusses broad reformulation and specific reformulations. For instance, Jansen, Zhang & Spink (2007) discussed three transition patterns, including the transition pattern between specialization and generalization. They found that 26.23% of queries showed users attempting to narrow the range of their results and 16.98% showed users attempting to generalize their results. Silverstein, Marais, Henzinger, and Moricz (1999) address two reasons for adding or subtracting terms from a query: the query is too specific or not specific enough.

Examining how users reformulate queries provides examples of how queries can
be placed on a continuum of specificity. Classifying the specificity of query reformulation requires comparisons of sequences of queries and the identification of each query as more or less specific than previous queries. My research aims to identify query specificity without this sequence of queries.

**Ambiguous Queries and Specificity**

Research has investigated identifying ambiguous queries and terms (Allan & Raghavan, 2002; Clough, Sanderson, Abouammoh, Navarro, & Paramita, 2009; Sanderson, 2008; Santos, Macdonald, & Ounis, 2010; Song, Luo, Nie, Yu, & Hon, 2009), which are sometimes linked to general queries. Ambiguous queries are queries that have multiple meanings or interpretations. In other words, ambiguous queries are queries for which the meaning is unclear. In terms of specificity, ambiguous queries cannot be identified as narrow without additional information. Thus ambiguous queries could be linked to my conception of general queries.

Ambiguous queries are often associated with short queries (Allan & Raghavan, 2002). According to Allan & Raghavan (2002) queries can be ambiguous for multiple reasons:

a) Because of the subject matter of the query  
b) Because of the user’s task is unclear  
c) Because the type of user is not known  
d) Because the type of information the user wants is unclear

Understanding these four aspects of a query is also important for identifying the
specificity of the query. Particularly, reasons b and d can be linked to the intent of a query and thus are important for Web search query specificity. As ambiguous queries and the specificity of queries are often times connected the research concerning ambiguous queries discussed below has implications for query specificity.

Researchers (Clough, et al., 2009; Santos, et al., 2010) have examined ambiguous queries and diversifying results for ambiguous queries. Diversifying results is one method for improving results for ambiguous queries. Sanderson (2008) explored ambiguous queries and the need for test collections that include ambiguous queries before describing a method for creating new collections from already existing resources and exploring the impact of query ambiguity on an information retrieval system.

Allan and Raghavan (2002) investigated how part-of-speech patterns can help reduce query ambiguity. They took short ambiguous queries and looked at how the query terms were related to the words near them in the document collection. The issue here is that their work assumes that you have the text of the documents in the overall collection; this cannot be assumed for Web searches.

Song, et al. (2009) explore identifying ambiguous queries, constructing a taxonomy of queries from ambiguous to specific. Their taxonomy has three levels (ambiguous, broad, and clear) and is of interest to my research as two of their types of queries are closely parallel to my general and narrow.

Examination of ambiguous queries often includes investigation of the usage of word sense disambiguation for queries (Allan & Raghavan, 2002). Word sense disambiguation looks at the different possible meanings of the terms in the query and tries to understand which meanings are correct for the query. This can be implicitly
inferrred in longer queries but is more difficult for short queries.

Related to ambiguous queries are the concepts of query clarity and query scope. Cronen-Townsend, Zhou, and Croft (2002) defined query clarity as a way to measure the coherence of the language usage in documents as compared to the query terms. Query clarity refers to the specificity/ambiguity of the query. He and Ounis (2004) studied the usage of several pre-retrieval predictors, including query clarity and query length. They also examined the idea of query scope, which is similar to query clarity, or the size of the document set that has at least one of the query terms. While both query clarity and query scope are related to the concept of Web search query specificity, unfortunately they again assume that one has the entire document collections with which to compare the queries. Thus they are not usable for classifying Web search queries without modification.

**Factors related to Specificity**

Web search query specificity is potentially influenced by many different factors. Most of these factors have not been examined directly in terms of their relationship to Web search query specificity. Potential factors include:

1. Query length
2. Part-of-speech usage in the query
3. Related terms to the query terms
4. Word sense disambiguation
5. Named entities in the query
6. Questions in the query
All five of these factors can potentially be useful for query specificity. However, current research usually looks at one term: query length. With ambiguous query research sometimes adding in other factors such as part-of-speech usage and word sense disambiguation. For our research, I focus on two factors of specificity: query length and parts of speech.

**Query Length**

Specificity is commonly associated with length where longer queries are considered to be more specific (Bailey, White, Liu, & Kumaran, 2010; Bendersky & Croft, 2009; R. Jones, Rey, Madani, & Greiner, 2006; Silverstein, et al., 1999). On the other hand, short queries are often associated with more ambiguous and general queries (Allan & Raghavan, 2002). While length is often cited as related to specificity of queries, few studies have empirically shown this. The exception being Phan, Bailey, and Wilkinson (2007), who used user studies to explore the relationship of information need specificity and search query length. They found that, on average, the crossover point for specificity of information need from narrow to broad is three words. Queries are more likely to be broader as the length decreases. However, query length is obviously not the only factor influencing specificity (González-Caro, et al., 2011; Phan, et al., 2007). In this research, I choose query length as a factor to explore for two reasons: 1) to test the commonly cited association between it and query specificity and 2) to serve as a benchmark for analysis of the usefulness of other factors.
**Part-of-Speech Usage**

An additional factor that I examine is the usage of parts of speech within individual queries. Research in computational linguistics and natural language processing has shown that techniques from these areas can help improve Web search outcomes (Kao & Poteet, 2004; Manshadi & Li, 2009). Analyses of part-of-speech usage show that search queries consist primarily of nouns (with common nouns and proper nouns constituting over 70% of the terms) and noun phrases (70% of a manually labeled sample of queries) (Barr, Jones, & Regelson, 2008). Similarly, Jansen, Spink and Pfaff (2000) find that most queries consist of noun phrases. Additionally, knowledge about the usage of parts of speech in queries has been employed for a variety of purposes such as to reduce query ambiguity (Allan & Raghavan, 2002) and to select words for query reformulation (Belkin, 1993). I analyze whether including information on part-of-speech usage can improve detection of the specificity of queries. For instance, if a query has a WH-Adverb (e.g. how, when, where, why, etc. used as an adverb) in it the implication is that it is a question, thus indicating a narrow intent. In this research, I choose part-of-speech usage as my second factor of web search query specificity to explore for three reasons: 1) it is automatically identified, 2) it has been previously linked to the related topic of ambiguous queries, and 3) there has been little research on how it could be used without document collections.
Other Studies related to Specificity

Other studies have looked explicitly at conceptions of specificity. For instance, Ramirez and Vries (2006) use the terms “broad” and “narrow” for the concepts of general and specific and investigate request type as a contextual feature. Rouet (2003) discusses task specificity and prior knowledge, particularly focusing on single versus multiple target questions. They differentiated between specific and general questions based on how much information needed to be located.

Lastly, whereas I use the term specificity to refer to the level of user intent behind the queries, other research has used the term differently. For instance in information retrieval, query specificity is used to refer to the sum and mean of Inverse Document Frequency (IDF) of query terms as in the research of Arampatzis and Kamps (2010) and K. S. Jones (1972). Related to this, Haiduc, Bavota, Oliveto, Marcus, and Lucia (2012) explored using a query specificity index to reflect the quality of a query. Their index is based on how common a term is over an entire document set. This use of the term specificity looks at the discriminative power of a query over a collection of documents and is distinct from the usage of the term in my research. It also has the drawback of applying only to individual terms and requiring prior knowledge of the overall document collection, which is not always feasible with Web searches.

Synthesis of Prior Work

Based on my review of related literature, I identified several trends. First, there are different ways to classify the intents and goals of queries. Second, while many of the
classifications of queries can be related to specificity, they do not directly address specificity across multiple topics and goals. They instead focus on small groups of queries that could be considered narrow or specific. Third, while work on the reformulation of queries often addresses the specialization or generalization of queries, it assumes multiple queries by a user on one topic. It does not address the specificity of queries that are not reformulated. Fourth, topics such as ambiguous queries and query clarity are closely related to specificity and have implications for identifying specificity but do not directly address the question of query specificity. Many of the methods for dealing with ambiguous queries and query clarity assume that the retrieval system has the full text of all possible documents and can examine them prior to retrieval. Fifth, identifying specificity of an information need is often times done manually or by users, not done algorithmically. Sixth, part-of-speech usage has been examined for Web search queries, but its ability to suggest specificity of query has not been explored.

Several open questions remain for research in query specificity, such as what factors other than length effect specificity? Can query length and other factors (e.g. parts of speech) be used to differentiate between narrow and general queries without a query sequence? How does part-of-speech usage relate to query specificity? Does using multiple factors improve identification of query specificity? Can users be separated into different groups based on part-of-speech usage?

In order to better understand the specificity of queries, we need to first understand what makes a query specific, what characteristics might indicate this across topics and users and how we can use these characteristics to measure the specificity of queries without requiring comparison to other related queries. In addition, we need a method to
automatically identify the specificity of queries without needing users’ judgment or the text of all of the potential returned webpages. These issues motivate my research, and I believe that operationalizing a definition of specificity across multiple users will be extremely helpful for understanding searching and improving search systems. I explore two factors of specificity using the commonly associated factor query length as a starting point and benchmark to evaluate a new factor, part of speech and the combination of both factors. By integrating these two factors, I hope to provide better classification of query specificity than either factor could singularly.
Chapter 3
Methodology

Dataset

I used a set of queries from a transaction log containing daily information about search queries. The transaction log contained 3.5 million searches from 65,000 users, from the AOL Search Service from March to May 2006. Each record in the transaction log contained information on the anonymous user, the date, time a query was submitted, the query, the type of search (e.g. web, images, etc.), and the query click URL.

I randomly selected 500 users from the larger dataset and then selected all queries in the dataset submitted by those users, resulting in a data set with 25,820 queries. I placed the dataset into three relational database tables linked by the query id and user id:

1) Query Table
2) Unique Query Table
3) User Table

The tables represent the information that I used during my research and the information that data was annotated with. Appendix A provides the names and descriptions of the dataset fields for each of the three tables.

The queries used 12,303 unique terms. On average, each unique query length was 3.06 terms and 19.10 characters. Among all of the queries (including duplicates across users), there is an average 2.86 terms per query and an average 17.71 characters per
query. On average, each user had 52.28 queries and 26.27 unique queries during the data collection period. Each user, on average, used 149.82 terms, 44.42 unique terms, 928.88 characters over all of their queries. Each user had an average number of 2.16 terms and 15.57 characters per query.

**Research Objective 1: Provide a systematic review of specificity to develop clear guidelines for identifying specificity**

I analyzed related work on specificity and types of queries. I reviewed the few articles that directly deal with query specificity and articles that examine the breadth of queries in order to investigate previously identified attributes or characteristics that could be related to query specificity. I also examined actual queries from the AOL Search Service transaction log based on query attributes found in the literature and through content analysis decided whether the attributes implied the level of query specificity. The examination of the queries was used to develop a list of nine attributes that more narrow queries contain.

**Research Objective 2: Evaluate the effectiveness of query length and parts of speech in identifying the specificity of queries**

For research objective 2, I have three steps:

1) Manually classify queries as narrow or general based on the nine attributes derived from research objective 1.

2) Calculate the query length of the queries in the dataset
3) Automatically identify the parts of speech present in the queries in the dataset.

4) Analyze the narrow and general queries based on query length and parts of speech and the combination of the two factors.

For step one, I randomly selected 5,115 unique queries from the overall AOL dataset and classified their specificity as narrow or general based on the list of nine attributes that I developed from research objective 1. The nine attributes represent narrow goals for queries and as such queries that had one or more of the attributes were labeled as narrow. All other queries were labeled as general. I realize that specificity is not binary. However, I reserve the probabilistic evaluation for future research.

For step two, I automatically calculated the number of terms and the number of characters in each query to represent query length. The average query in the dataset has 3.56 terms and 21.27 characters.

For step three to determine part-of-speech usage, I used a part-of-speech identifier tool from Stanford Natural Language Processing Group (Toutanova, Klen, Manning, & Singer, 2003; Toutanova & Manning, 2000). To identify the parts of speech for the words in a given text, I constructed a Java program that interfaces with the identifier. The part of speech identifier used the Penn Treebank tag set to label parts of speech (Marcus, Santorini, & Marcinkiewicz, 1993). I used these same labels for the subcategories of parts of speech. Appendix B gives the part of speech tags and their meanings.

I counted the usage of five broad categories of parts of speech: nouns, adjectives, adverbs, verbs and other. I also looked at subcategories of parts of speech, such as the type of nouns (e.g., proper noun, singular or noun, singular or mass) in the broad category of nouns (See Table 3-1). Appendix C gives an example of a query containing the most
common parts of speech used in the dataset. Note that for some of the rarer parts of speech I do not give query examples. I analyzed both the broad categories and the subcategories. For the unique queries, 95% had a noun, 27% had adjectives, 7% had adverbs, 26% had verbs and 36% had other parts of speech. I excluded from my analysis any tagged part of speech that was present in fewer than 25 queries in the sample (0.5% of dataset) due to the small sample size. The overall results are similar to those found by Jansen, et al. (2000).

<table>
<thead>
<tr>
<th>Broad Categories</th>
<th>Subcategories</th>
<th>Frequency in dataset (# of queries present)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Nouns</td>
<td>Noun, singular or mass</td>
<td>4,864</td>
</tr>
<tr>
<td></td>
<td>Noun, plural</td>
<td>4,358</td>
</tr>
<tr>
<td></td>
<td>Proper Noun, singular</td>
<td>1858</td>
</tr>
<tr>
<td></td>
<td>Proper Noun, plural</td>
<td>26</td>
</tr>
<tr>
<td>Adjectives</td>
<td>Adjectives</td>
<td>1,364</td>
</tr>
<tr>
<td></td>
<td>Adjectives, comparative</td>
<td>1,297</td>
</tr>
<tr>
<td></td>
<td>Adjectives, superlative</td>
<td>36</td>
</tr>
<tr>
<td></td>
<td>Adverbs</td>
<td>372</td>
</tr>
<tr>
<td></td>
<td>Adverbs, comparative</td>
<td>283</td>
</tr>
<tr>
<td></td>
<td>Adverbs, superlative</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>Wh-Adverbs</td>
<td>98</td>
</tr>
<tr>
<td>Verbs</td>
<td>Verb, base form</td>
<td>1,314</td>
</tr>
<tr>
<td></td>
<td>Verb, past tense</td>
<td>313</td>
</tr>
<tr>
<td></td>
<td>Verb, gerund or present participle</td>
<td>151</td>
</tr>
<tr>
<td></td>
<td>Verb, past participle</td>
<td>242</td>
</tr>
<tr>
<td></td>
<td>Verb, non-3rd person singular present</td>
<td>137</td>
</tr>
<tr>
<td></td>
<td>Verb, 3rd person singular present</td>
<td>398</td>
</tr>
<tr>
<td></td>
<td>Coordinating conjunctions</td>
<td>248</td>
</tr>
<tr>
<td></td>
<td>Other</td>
<td>1,831</td>
</tr>
<tr>
<td></td>
<td></td>
<td>155</td>
</tr>
</tbody>
</table>
Table 3-1: The different parts of speech automatically identified and their frequency in the dataset

<table>
<thead>
<tr>
<th>Broad Categories</th>
<th>Subcategories</th>
<th>Frequency in data set (# of queries present)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cardinal numbers</td>
<td></td>
<td>398</td>
</tr>
<tr>
<td>Determiners</td>
<td></td>
<td>399</td>
</tr>
<tr>
<td>Existential there</td>
<td></td>
<td>0</td>
</tr>
<tr>
<td>Foreign word</td>
<td></td>
<td>179</td>
</tr>
<tr>
<td>Preposition or subordinating conjunction</td>
<td></td>
<td>879</td>
</tr>
<tr>
<td>List item marker</td>
<td></td>
<td>20</td>
</tr>
<tr>
<td>Modal</td>
<td></td>
<td>20</td>
</tr>
<tr>
<td>Predeterminer</td>
<td></td>
<td>3</td>
</tr>
<tr>
<td>Possessive ending</td>
<td></td>
<td>0</td>
</tr>
<tr>
<td>Personal pronoun</td>
<td></td>
<td>170</td>
</tr>
<tr>
<td>Possessive pronoun</td>
<td></td>
<td>107</td>
</tr>
<tr>
<td>Particle</td>
<td></td>
<td>37</td>
</tr>
<tr>
<td>Symbol</td>
<td></td>
<td>51</td>
</tr>
<tr>
<td>To</td>
<td></td>
<td>152</td>
</tr>
<tr>
<td>Interjection</td>
<td></td>
<td>11</td>
</tr>
<tr>
<td>Wh-determiner</td>
<td></td>
<td>21</td>
</tr>
<tr>
<td>Wh-pronoun</td>
<td></td>
<td>72</td>
</tr>
<tr>
<td>Possessive wh-pronoun</td>
<td></td>
<td>0</td>
</tr>
</tbody>
</table>

Finally, in order to double check the part-of-speech identifier, I submitted the queries to an additional freely available tagger TreeTagger (Schmid, 1994, 1995), which also uses the Penn Treebank tag set to label parts of speech. I compared the percentage usage for unique queries for each part of speech tagger. Of the 5 broad categories and 36 subcategories, 25 had a difference in percentage of less than 1%, 13 had a difference in usage of between 1-3%, and 3 [singular proper noun, adjective (broad category) and adjective (subcategory)] had a change between 8-12%.

For step four, I analyzed the manually classified unique queries on three levels. First, I calculated the average query length for narrow and general queries and performed an unpaired t-test to evaluate whether the differences between narrow and general queries
were statistically significant. Second, I examined the percentage of queries containing each part of speech classified as narrow and general to evaluate whether the part of speech gave an indication of query specificity. Here I performed Chi-squares tests to evaluate whether the differences were statistically significant. Third, I explored whether combining query length and parts of speech improved identification of query specificity. Again here I performed Chi-squares tests to evaluate whether the differences were statistically significant.

**Research Objective 3: Automatically identify queries as narrow or general based on results of an analysis of query length and parts of speech for manually classified queries**

Based on the analysis from research objective 2, I developed five algorithms for classifying queries as either narrow or general. One algorithm, Algorithm 1, used only query length. Two, Algorithms 2 and 3, used only parts of speech. The difference between algorithms 2 and 3 was that they used different selections of parts of speech for their identification of specificity. Algorithm 2 used the parts of speech: noun, singular proper; wh-adverb; verb, base form; preposition or subordinate conjunction; particle; to; wh-pronoun. Algorithm 3 used the parts of speech: adverb; verb; other; noun, singular proper; adjectives, comparative; adjectives, superlative. The final two algorithms used both query length and parts of speech, combined in different manners. Algorithm 4 uses Algorithm 1 and 2. Algorithm 5 uses Algorithm 1 and 3. The algorithms were encoded in Java code and used to label the 25,820 Web search queries in the dataset. To evaluate the accuracy of the algorithms, I also compared the classification for the 5,115 queries I
manually labeled to the same queries in the algorithms’ output. The process of classifying a set of queries was as follows:

1) Load queries into a relational database.

2) Calculate the number of terms and number of characters in each query.

3) Identify the parts of speech in each query.

4) Submit the query, the number of terms, the number of characters and the parts of speech to the chosen algorithm.

**Research Objective 4: Explore whether I can automatically identify distinct user groups by their patterns of part-of-speech-usage**

Here I examined statistics generated from users’ use of different parts of speech in their queries. The sequence for this objective was as follows:

1) Load a sample into a relational database.

2) Identify the parts of speech for each unique query.

3) Count the frequencies of the parts of speech for each user.

4) Analyze the descriptive statistics of the unique queries for part-of-speech usage.

5) Attempt to separate users into distinct groups with K-Means Clustering based on the parts of speech used.

6) Analyze the different clusters of users to see the relationship among the groups and specificity of Web search queries.

The first task for this research objective was to represent the presence of particular parts of speech in a way that makes the different users comparable. Due to the large number of parts of speech and the infrequency of some of the parts of speech I used
only the broad categories of parts of speech (noun, adjective, adverb, verb and other) for this portion of the research. The first step was to transform the count data for parts of speech so that they were comparable across users. Otherwise, it would have been difficult to compare the number of each part of speech that different users used because this was highly dependent on the number of queries a user entered. So, a user who entered more queries would have higher values for many of the counts than a user with fewer queries. I created two transformed variables from the count data. The first was the average number of the parts of speech per query and the second was the percentage of terms entered by the user that corresponded to the part of speech. I examined which variable might be most useful and decided on the percentage of terms because it provided a standard range of values (0-1) for every user and was less likely to be confounded with query length than the average number per query.

K-means clustering aims to partition the points of data (in my case users) into k groups so that the sum of squares from the data points to the assigned cluster centers is minimized. I used an implementation of Hartigan and Wong (1979)’s algorithm in the R environment (Team, 2012). After applying K-means clustering to separate the users into distinct groups the clusters are examined in terms of the percentage of terms that fall into each broad part of speech category. To check whether differences were statistically significant we used Wilcoxon tests to compare the clusters by each broad part of speech category.
Chapter 4
Results

Research Objective 1: Provide a systematic review of specificity to develop clear guidelines for identifying specificity

While specificity is an important aspect of search queries, (González-Caro & Baeza-Yates, 2011; González-Caro, et al., 2011) it is difficult to measure. Specificity is also difficult to define because it is relative and varies depending on the user and the topic. As such, many researchers deal with limited aspects of specificity without giving a full definition of the concept. Here I divide research on specificity into two distinct categories: research addressing partial aspects of the concept and research addressing the whole concept.

Research that has dealt with subsets of the concept identify queries that have a particular narrow goal or topic such as product queries (Shen, et al., 2009), math queries (Kamali, et al., 2012), geographic queries (R. Jones, et al., 2008), software queries (U. Lee, Liu, & Cho, 2005), person name queries (U. Lee, et al., 2005), or navigational queries (W. M. Lee & Sanderson, 2010). These works provide important information on narrow queries, but they do not define the larger concept of specificity.

Despite research on subsets of narrow queries, few researchers have examined the overall concept of specificity. Table 4-1 summarizes definitions and associations in the literature. Definitions of specificity in the literature range from vague (Phan, et al., 2007; Resnick & Vaughan, 2006) to more detailed (González-Caro & Baeza-Yates, 2011; González-Caro, et al., 2011; Rouet, 2003). Often longer queries are described as being
more specific or having more specific keywords; however, discussion of the meaning of “more specific” or “more general” is rare. Rouet (2003) examines task specificity and directly looks at question specificity and compares single versus multiple target questions. Task specificity is often a part of query specificity.

Table 4-1: Definitions and associations with specificity

<table>
<thead>
<tr>
<th>Definition/Association</th>
<th>Research</th>
</tr>
</thead>
<tbody>
<tr>
<td>Associated with length</td>
<td>(Bailey, et al., 2010; Bendersky &amp; Croft, 2009; R. Jones, et al., 2006; Silverstein, et al., 1999)</td>
</tr>
<tr>
<td>Longer implies more Narrow/Shorter implies more General</td>
<td></td>
</tr>
<tr>
<td>Specific/General Keywords (without definition)</td>
<td>(Resnick &amp; Vaughan, 2006)</td>
</tr>
<tr>
<td>More specific information needs (without definition)</td>
<td>(Resnick &amp; Vaughan, 2006)</td>
</tr>
<tr>
<td>Narrow versus broad</td>
<td>(Phan, et al., 2007)</td>
</tr>
<tr>
<td>Indicate there are factors other than length</td>
<td>(Phan, et al., 2007)</td>
</tr>
<tr>
<td>Different types/goals of questions- what kind of information needed/how to get the information</td>
<td>(Rouet, 2003)</td>
</tr>
<tr>
<td>Three levels (Specific, Medium, Broad)</td>
<td>(González-Caro, et al., 2011)</td>
</tr>
<tr>
<td>Three Levels (Ambiguous, Broad, Clear)</td>
<td>(Song, et al., 2009)</td>
</tr>
</tbody>
</table>

The most detailed description of specificity for queries comes from Gonzalez-Caro et al. (2011) who divide queries into three groups: specific, medium and broad, with specific queries being those that have a name, date, place, acronym or URL. Medium being those queries that are more general, and broad are those queries that have a very general term. Another clear taxonomy comes from Song, et al. (2009) who classify queries as:

Type A (ambiguous query): A query with multiple meanings

Type B (Broad query): a query that covers many subtopics

Type C (Clear query): a query that covers a subtopic.
The question that arises from this classification is how does one decide what is a broad topic and what is a subtopic?

González-Caro, et al. (2011) connect specificity to user intent. Similarly, Rouet (2003) links specificity and tasks and Song, et al. (2009) link specificity to the type of topics a query covers. These three ideas are similar in that understanding the specificity of a task can aid in identifying the user intent behind a query. Additionally, understanding the topics of a query can aid in detecting the user intent behind a query. In turn, distinguishing the user intent of a query leads to a better understanding of the specificity of the query. As such, for this research, I define query specificity as how narrow or general the user intent behind the query is. Jansen, et al. (2008) present a detailed three level taxonomy of user intent. The highest level is informational, navigational and transactional queries. Level two and level three provide more granular classifications, which inform my choice of attributes. (e.g. Level two: Informational, List; Informational, Advice; Level three: Informational, Directed, Closed; Informational, Directed, Open; Transactional, Results, Links).

The next step was to develop a list of attributes that queries with narrow user intent contain. I analyzed queries from the dataset and selected nine attributes to identify specificity. The queries where examined using content analysis for previous identified attributes or characteristics of queries that could be related to query specificity and other attributes that could be clearly identified from analysis of the queries. Because the intent behind the query is important for understanding specificity, I selected attributes that indicate the intent of the user or that have been identified by past researchers as qualities of narrow queries. The nine attributes are all connected to narrow user intents. For
example, if a query contains a URL, the user wishes to reach a specific website or if a query contains a request for directions, then a user is looking for a list of steps to follow to reach a goal. Table 4-2 shows the nine attributes I chose with example queries.

<table>
<thead>
<tr>
<th>Attribute</th>
<th>Example Query</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Query contains a URL</td>
</tr>
<tr>
<td>2</td>
<td>Query contains a location or name along with additional terms</td>
</tr>
<tr>
<td>3</td>
<td>Query compares multiple things</td>
</tr>
<tr>
<td>4</td>
<td>Query contains multiple distinct ideas or topics</td>
</tr>
<tr>
<td>5</td>
<td>Query contains a question that has a clear answer</td>
</tr>
<tr>
<td>6</td>
<td>Query contains a request for directions, instructions, tips</td>
</tr>
<tr>
<td>7</td>
<td>Query contains a specific date and additional terms</td>
</tr>
<tr>
<td>8</td>
<td>Query contains a number and additional terms</td>
</tr>
<tr>
<td>9</td>
<td>Query contains a name and additional terms</td>
</tr>
</tbody>
</table>

Based on these attributes, I defined query specificity at two levels narrow and general. Narrow queries were those that had one or more of the above attributes. General queries were those that did not contain any of the attributes.

**Research Objective 2: Evaluate the effectiveness of query length and parts of speech in identifying the specificity of queries**

Based on the attributes, I classified 5,115 unique queries from the larger dataset as either narrow or general: 62% (3,103) of the queries as narrow and 38% (2,012) of the queries as general. Next I analyzed the queries in terms of query length and part-of-
speech usage.

**Query Length**

On average, narrow queries were twice as long as general queries in terms of the number of terms and the number of characters they contained (unpaired t-test p<0.0001) (Table 4-3).

<table>
<thead>
<tr>
<th></th>
<th>Narrow</th>
<th>General</th>
</tr>
</thead>
<tbody>
<tr>
<td># Unique Queries</td>
<td>3,103</td>
<td>2,012</td>
</tr>
<tr>
<td>Average # Terms</td>
<td>4.51</td>
<td>2.1</td>
</tr>
<tr>
<td>Average # Characters</td>
<td>26.6</td>
<td>13.06</td>
</tr>
</tbody>
</table>

Examining the queries in more detail (Table 4-4), I found that increasing the number of terms improves accuracy of determining specificity. Only 3% of queries with 4 or more terms are classified as general, whereas 97% are classified as narrow. A similar pattern is shown for the number of characters in a query with the number of general queries decreasing as the number of characters in a query increases (Table 4-4). Only 12% of queries with 20 or more characters are classified as general, as compared to 88% classified as narrow. These findings on the relationship between specificity and query length are statistically significant at the p<0.0001 level (Table 4-4, Table 4-5).
Table 4-4: Effect of increasing the number of terms on the percentage of narrow and general queries classified. Chi square for number of terms * p<0.0001.

<table>
<thead>
<tr>
<th>Query Length (# of Terms in Query)</th>
<th>% of Queries Classified as Narrow</th>
<th>% of Queries Classified as General</th>
<th>$\chi^2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>&gt;=1 Term</td>
<td>62%</td>
<td>38%</td>
<td>--</td>
</tr>
<tr>
<td>&gt;=2 Terms</td>
<td>68%</td>
<td>32%</td>
<td>68.8*</td>
</tr>
<tr>
<td>&gt;=3 Terms</td>
<td>82%</td>
<td>18%</td>
<td>546.1*</td>
</tr>
<tr>
<td>&gt;=4 Terms</td>
<td>93%</td>
<td>7%</td>
<td>779.1*</td>
</tr>
<tr>
<td>&gt;=5 Terms</td>
<td>97%</td>
<td>3%</td>
<td>608.5*</td>
</tr>
</tbody>
</table>

Table 4-5: Effect of increasing number of characters on the percentage of narrow and general queries classified. Chi square for number of characters * p<0.0001, N.S. not significant.

<table>
<thead>
<tr>
<th>Query Length (# of Characters in Query)</th>
<th>% of Queries Classified as Narrow</th>
<th>% of Queries Classified as General</th>
<th>$\chi^2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>&gt;=1 Characters</td>
<td>62%</td>
<td>38%</td>
<td>--</td>
</tr>
<tr>
<td>&gt;=5 Characters</td>
<td>62%</td>
<td>38%</td>
<td>0.2 N.S.</td>
</tr>
<tr>
<td>&gt;=10 Characters</td>
<td>68%</td>
<td>32%</td>
<td>68.0*</td>
</tr>
<tr>
<td>&gt;=15 Characters</td>
<td>78%</td>
<td>22%</td>
<td>352.7*</td>
</tr>
<tr>
<td>&gt;=20 Characters</td>
<td>88%</td>
<td>12%</td>
<td>684.9*</td>
</tr>
<tr>
<td>&gt;=25 Characters</td>
<td>95%</td>
<td>5%</td>
<td>704.0*</td>
</tr>
</tbody>
</table>

Part-of-Speech Usage

For part-of-speech usage, I found that for broad categories (Table 4-6) queries containing adverbs, verbs and other parts of speech are more likely to be narrow (77% classified as narrow) than are queries with nouns and adjectives (63% and 67%, respectively). All of these groups, except nouns, are statistically significant (Table 4-6).
Table 4-6: Percentages of queries with each broad part of speech that are classified as narrow. Chi square for broad parts of speech * p<0.0001, N.S. not significant.

<table>
<thead>
<tr>
<th>Broad Part of Speech</th>
<th>% queries classified as narrow</th>
<th>$\chi^2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Nouns</td>
<td>63%</td>
<td>2.9 N.S.</td>
</tr>
<tr>
<td>Adjectives</td>
<td>67%</td>
<td>16.3*</td>
</tr>
<tr>
<td>Adverbs</td>
<td>77%</td>
<td>37.5*</td>
</tr>
<tr>
<td>Verbs</td>
<td>77%</td>
<td>133.5*</td>
</tr>
<tr>
<td>Other</td>
<td>77%</td>
<td>166.2*</td>
</tr>
</tbody>
</table>

Including information about the attributes of various parts of speech used in queries aids in the detection of specificity resulting in percentages in the 70%-90% range (Table 4-7). The exception to this is the category symbols, which is as likely to be general as narrow (51% classified as narrow.) Most of these relationships are statistically significant except for comparative adjectives and symbols (Table 4-7). In the case of comparative adjectives, the lack of significance may be due to the small sample size.

Table 4-7: Percentages of queries with each part of speech that are classified as narrow. Chi square for parts of speech * p<0.0001, **p<0.01, ***p<0.05, N.S. not significant.

<table>
<thead>
<tr>
<th>Part of Speech</th>
<th>% queries classified as narrow</th>
<th>$\chi^2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>All Queries</td>
<td>62%</td>
<td>--</td>
</tr>
<tr>
<td>No Nouns</td>
<td>12%</td>
<td>266.9*</td>
</tr>
<tr>
<td>Nouns, Singular or Mass</td>
<td>68%</td>
<td>42.6*</td>
</tr>
<tr>
<td>Nouns, Plural</td>
<td>66%</td>
<td>13.2**</td>
</tr>
<tr>
<td>Nouns, Proper</td>
<td>85%</td>
<td>5.6***</td>
</tr>
<tr>
<td>Adjectives</td>
<td>67%</td>
<td>14.2**</td>
</tr>
<tr>
<td>Adjectives, Comparative</td>
<td>75%</td>
<td>2.6 N.S.</td>
</tr>
<tr>
<td>Adjectives, Superlative</td>
<td>76%</td>
<td>3.9***</td>
</tr>
<tr>
<td>Adverb</td>
<td>72%</td>
<td>11.4**</td>
</tr>
<tr>
<td>Wh-Adverbs</td>
<td>96%</td>
<td>47.9*</td>
</tr>
<tr>
<td>Verb, Base form</td>
<td>84%</td>
<td>62.6*</td>
</tr>
<tr>
<td>Verb, Past Tense</td>
<td>77%</td>
<td>14.1**</td>
</tr>
<tr>
<td>Verb, Gerund or Present Participle</td>
<td>72%</td>
<td>10.9**</td>
</tr>
<tr>
<td>Verb, Past Participle</td>
<td>72%</td>
<td>6.1**</td>
</tr>
</tbody>
</table>
Table 4-7: Percentages of queries with each part of speech that are classified as narrow. Chi square for parts of speech * p<0.0001, **p<0.01, ***p<0.05, N.S. not significant.

<table>
<thead>
<tr>
<th>Part of Speech</th>
<th>% queries classified as narrow</th>
<th>( \chi^2 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>Verb, Non-3rd Person Singular Present</td>
<td>81%</td>
<td>63.6*</td>
</tr>
<tr>
<td>Verb, 3rd Person Singular Present</td>
<td>84%</td>
<td>52.2*</td>
</tr>
<tr>
<td>Coordinating Conjunction</td>
<td>82%</td>
<td>26.1*</td>
</tr>
<tr>
<td>Cardinal Numbers</td>
<td>81%</td>
<td>58.8*</td>
</tr>
<tr>
<td>Determiners</td>
<td>78%</td>
<td>45.9*</td>
</tr>
<tr>
<td>Foreign Words</td>
<td>70%</td>
<td>4.7***</td>
</tr>
<tr>
<td>Preposition or Subordinate Conjunction</td>
<td>86%</td>
<td>211.0*</td>
</tr>
<tr>
<td>Personal Pronoun</td>
<td>76%</td>
<td>15.1*</td>
</tr>
<tr>
<td>Possessive Pronoun</td>
<td>74%</td>
<td>6.4***</td>
</tr>
<tr>
<td>Particle</td>
<td>84%</td>
<td>7.5***</td>
</tr>
<tr>
<td>Symbol</td>
<td>51%</td>
<td>2.6 N.S.</td>
</tr>
<tr>
<td>to</td>
<td>91%</td>
<td>55.9*</td>
</tr>
<tr>
<td>Wh-pronoun</td>
<td>99%</td>
<td>41.0*</td>
</tr>
</tbody>
</table>

Finally, I evaluated whether combining information about parts of speech and query length (Table 4-8, Table 4-9) improves identification of narrow queries. For all broad parts of speech if a query is three or more terms there was a greater than 80% chance that it is narrow; all of these findings are statistically significant (Table 4-8). Similarly, for subcategories of parts of speech, having three or more terms increased the likelihood that a query is narrow (Table 4-9). Most of these relationships are statistically significant; in the cases where they are not this may be due to small sample size (Table 4-9). Interestingly, the increase in the percentage of narrow queries was often greater when moving from two or more terms to three or more terms than from one or more terms to two or more terms. For example, the broad part of speech nouns increased from 63% to 69% for one or more terms to two or more terms and then from 69% to 83% when going
to three or more terms.

Table 4-8: Percentages of queries with each part of speech that are classified as narrow. Chi square for parts of speech * p<0.0001, **p<0.01, ***p<0.05, N.S. not significant.

<table>
<thead>
<tr>
<th>Broad Part of Speech</th>
<th>&gt;= 1 Term in Query</th>
<th>&gt;=2 Terms in Query</th>
<th>&gt;=3 Terms in Query</th>
<th>χ²</th>
</tr>
</thead>
<tbody>
<tr>
<td>Nouns</td>
<td>63%</td>
<td>2.9 N.S.</td>
<td>69%</td>
<td>95.7*</td>
</tr>
<tr>
<td>Adjectives</td>
<td>67%</td>
<td>16.3*</td>
<td>70%</td>
<td>34.5*</td>
</tr>
<tr>
<td>Adverbs</td>
<td>77%</td>
<td>37.5*</td>
<td>78%</td>
<td>42.3*</td>
</tr>
<tr>
<td>Verbs</td>
<td>77%</td>
<td>133.5*</td>
<td>79%</td>
<td>159.7*</td>
</tr>
<tr>
<td>Other</td>
<td>77%</td>
<td>166.2*</td>
<td>79%</td>
<td>210.7*</td>
</tr>
</tbody>
</table>

Table 4-9: Percentages of queries with each part of speech that are classified as narrow. – sample size <25. Chi square for parts of speech * p<0.0001, **p<0.01, ***p<0.05, N.S. not significant.

<table>
<thead>
<tr>
<th>Part of Speech</th>
<th>&gt;= 1 Term in Query</th>
<th>&gt;=2 Terms in Query</th>
<th>&gt;=3 Terms in Query</th>
<th>χ²</th>
</tr>
</thead>
<tbody>
<tr>
<td>No Nouns</td>
<td>12%</td>
<td>266.9*</td>
<td>25%</td>
<td>69.7*</td>
</tr>
<tr>
<td>Nouns, Singular or Mass</td>
<td>68%</td>
<td>42.6*</td>
<td>72%</td>
<td>171.2*</td>
</tr>
<tr>
<td>Nouns, Plural</td>
<td>66%</td>
<td>13.2**</td>
<td>70%</td>
<td>48.2*</td>
</tr>
<tr>
<td>Nouns, Proper</td>
<td>85%</td>
<td>5.6***</td>
<td>85%</td>
<td>5.6***</td>
</tr>
<tr>
<td>Adjectives</td>
<td>67%</td>
<td>14.2**</td>
<td>69%</td>
<td>29.8*</td>
</tr>
<tr>
<td>Adjectives, Comparative</td>
<td>75%</td>
<td>2.6 N.S.</td>
<td>79%</td>
<td>4.4***</td>
</tr>
<tr>
<td>Adjectives, Superlative</td>
<td>76%</td>
<td>3.9***</td>
<td>81%</td>
<td>6.9**</td>
</tr>
<tr>
<td>Adverb</td>
<td>72%</td>
<td>11.4**</td>
<td>73%</td>
<td>14.3**</td>
</tr>
<tr>
<td>Wh-Adverbs</td>
<td>96%</td>
<td>47.9*</td>
<td>96%</td>
<td>47.9*</td>
</tr>
<tr>
<td>Verb, Base form</td>
<td>84%</td>
<td>62.6*</td>
<td>85%</td>
<td>71.0*</td>
</tr>
<tr>
<td>Verb, Past Tense</td>
<td>77%</td>
<td>14.1**</td>
<td>79%</td>
<td>17.8*</td>
</tr>
<tr>
<td>Verb, Gerund or Present Participle</td>
<td>72%</td>
<td>10.9**</td>
<td>75%</td>
<td>17.0*</td>
</tr>
</tbody>
</table>
Table 4-9: Percentages of queries with each part of speech that are classified as narrow. – sample size <25. Chi square for parts of speech * p<0.0001, **p<0.01, ***p<0.05, N.S. not significant.

<table>
<thead>
<tr>
<th>Part of Speech</th>
<th>&gt;= 1 Term in Query</th>
<th>( \chi^2 )</th>
<th>&gt;=2 Terms in Query</th>
<th>( \chi^2 )</th>
<th>&gt;=3 Terms in Query</th>
<th>( \chi^2 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>Verb, Past Participle</td>
<td>72%</td>
<td>6.1**</td>
<td>74%</td>
<td>8.7***</td>
<td>83%</td>
<td>22.2*</td>
</tr>
<tr>
<td>Verb, Non-3rd Person Singular Present</td>
<td>81%</td>
<td>63.6*</td>
<td>82%</td>
<td>67.2*</td>
<td>90%</td>
<td>114.8*</td>
</tr>
<tr>
<td>Verb, 3rd Person Singular Present</td>
<td>84%</td>
<td>52.2*</td>
<td>85%</td>
<td>53.6*</td>
<td>89%</td>
<td>71.8*</td>
</tr>
<tr>
<td>Coordinating Conjunction</td>
<td>82%</td>
<td>26.1*</td>
<td>82%</td>
<td>26.1*</td>
<td>82%</td>
<td>26.1*</td>
</tr>
<tr>
<td>Cardinal Numbers</td>
<td>81%</td>
<td>58.8*</td>
<td>84%</td>
<td>76.0*</td>
<td>90%</td>
<td>113.8*</td>
</tr>
<tr>
<td>Determiners</td>
<td>78%</td>
<td>45.9*</td>
<td>79%</td>
<td>48.8*</td>
<td>81%</td>
<td>59.7*</td>
</tr>
<tr>
<td>Foreign Words</td>
<td>70%</td>
<td>4.7***</td>
<td>70%</td>
<td>5.1***</td>
<td>81%</td>
<td>23.1*</td>
</tr>
<tr>
<td>Preposition or Subordinate Conjunction</td>
<td>86%</td>
<td>211.0*</td>
<td>88%</td>
<td>247.2*</td>
<td>89%</td>
<td>252.8*</td>
</tr>
<tr>
<td>Personal Pronoun</td>
<td>76%</td>
<td>15.1*</td>
<td>77%</td>
<td>16.9*</td>
<td>83%</td>
<td>27.9*</td>
</tr>
<tr>
<td>Possessive Pronoun</td>
<td>74%</td>
<td>6.4***</td>
<td>75%</td>
<td>7.1***</td>
<td>83%</td>
<td>18.1*</td>
</tr>
<tr>
<td>Particle</td>
<td>84%</td>
<td>7.5***</td>
<td>84%</td>
<td>7.5***</td>
<td>88%</td>
<td>10.5**</td>
</tr>
<tr>
<td>Symbol</td>
<td>51%</td>
<td>2.6 N.S.</td>
<td>55%</td>
<td>0.9 N.S.</td>
<td>66%</td>
<td>0.2 N.S.</td>
</tr>
<tr>
<td>To</td>
<td>91%</td>
<td>55.9*</td>
<td>91%</td>
<td>55.9*</td>
<td>91%</td>
<td>55.4*</td>
</tr>
<tr>
<td>Wh-pronoun</td>
<td>99%</td>
<td>41.0*</td>
<td>99%</td>
<td>41.0*</td>
<td>99%</td>
<td>41.0*</td>
</tr>
</tbody>
</table>

Research Objective 3: Automatically identify queries as narrow or general based on results of an analysis of query length and parts of speech for manually classified queries

I automatically labeled the entire dataset of 25,820 Web search queries with their corresponding query lengths and parts of speech. Nouns were the most common broad
part of speech, but the specific forms of nouns were less common (Table 4-10). The other broad parts of speech also occurred in queries. A few subcategories (e.g. existential there, possessive endings, and possessive wh-pronouns) were not present in any query. Most likely these parts of speech are rare in queries even in a larger sample.

Queries contained many different combinations of parts of speech such as noun adjective or noun adjective verb (Table 4-10). Most of the combinations were present in the queries. On average, the more parts of speech in a query, the longer it was.

<table>
<thead>
<tr>
<th>Table 4-10: Statistics for part of speech usage in unique queries</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Grouping</strong></td>
</tr>
<tr>
<td><strong>Used Specific Part of Speech</strong></td>
</tr>
<tr>
<td>Without a Noun</td>
</tr>
<tr>
<td>With Something Other Than a Noun</td>
</tr>
<tr>
<td>With Only Nouns</td>
</tr>
<tr>
<td>Noun</td>
</tr>
<tr>
<td>Noun, singular or mass</td>
</tr>
<tr>
<td>Noun, plural</td>
</tr>
<tr>
<td>Proper Noun, singular</td>
</tr>
<tr>
<td>Proper Noun, plural</td>
</tr>
<tr>
<td>Adjective</td>
</tr>
<tr>
<td>Adjective, comparative</td>
</tr>
<tr>
<td>Adjective, superlative</td>
</tr>
<tr>
<td>Adverb</td>
</tr>
<tr>
<td>Adverb, comparative</td>
</tr>
<tr>
<td>Adverb, superlative</td>
</tr>
<tr>
<td>Wh-Adverbs</td>
</tr>
<tr>
<td>Verb</td>
</tr>
<tr>
<td>Verb, base form</td>
</tr>
<tr>
<td>Verb, past tense</td>
</tr>
<tr>
<td>Verb, gerund or present participle</td>
</tr>
</tbody>
</table>
Table 4-10: Statistics for part of speech usage in unique queries

<table>
<thead>
<tr>
<th>Grouping</th>
<th># Unique Queries</th>
<th>% Unique Queries</th>
<th>Average # Terms</th>
<th>Average # Characters</th>
</tr>
</thead>
<tbody>
<tr>
<td>Verb, past participle</td>
<td>137</td>
<td>1%</td>
<td>6.39</td>
<td>37.50</td>
</tr>
<tr>
<td>Verb, non-3rd person singular present</td>
<td>571</td>
<td>5%</td>
<td>5.37</td>
<td>30.64</td>
</tr>
<tr>
<td>Verb, 3rd person singular present</td>
<td>252</td>
<td>2%</td>
<td>6.40</td>
<td>36.15</td>
</tr>
<tr>
<td>Other</td>
<td>2,715</td>
<td>23%</td>
<td>5.02</td>
<td>27.38</td>
</tr>
<tr>
<td>Coordinating conjunctions</td>
<td>213</td>
<td>2%</td>
<td>7.86</td>
<td>44.56</td>
</tr>
<tr>
<td>Cardinal numbers</td>
<td>634</td>
<td>5%</td>
<td>5.16</td>
<td>28.03</td>
</tr>
<tr>
<td>Determiners</td>
<td>522</td>
<td>4%</td>
<td>6.51</td>
<td>33.97</td>
</tr>
<tr>
<td>Existential there</td>
<td>0</td>
<td>&lt;1%</td>
<td>--</td>
<td>--</td>
</tr>
<tr>
<td>Foreign word</td>
<td>179</td>
<td>1%</td>
<td>6.05</td>
<td>27.94</td>
</tr>
<tr>
<td>Preposition or subordinating conjunction</td>
<td>1,453</td>
<td>12%</td>
<td>5.69</td>
<td>31.60</td>
</tr>
<tr>
<td>List item marker</td>
<td>20</td>
<td>&lt;1%</td>
<td>3.0</td>
<td>12.55</td>
</tr>
<tr>
<td>Modal</td>
<td>20</td>
<td>&lt;1%</td>
<td>3.0</td>
<td>12.55</td>
</tr>
<tr>
<td>Predeterminer</td>
<td>3</td>
<td>&lt;1%</td>
<td>12.00</td>
<td>57.00</td>
</tr>
<tr>
<td>Possessive ending</td>
<td>0</td>
<td>&lt;1%</td>
<td>--</td>
<td>--</td>
</tr>
<tr>
<td>Personal pronoun</td>
<td>170</td>
<td>1%</td>
<td>7.84</td>
<td>38.66</td>
</tr>
<tr>
<td>Possessive pronoun</td>
<td>107</td>
<td>1%</td>
<td>8.29</td>
<td>42.79</td>
</tr>
<tr>
<td>Particle</td>
<td>37</td>
<td>&lt;1%</td>
<td>7.49</td>
<td>36.86</td>
</tr>
<tr>
<td>Symbol</td>
<td>51</td>
<td>&lt;1%</td>
<td>5.57</td>
<td>23.65</td>
</tr>
<tr>
<td>To</td>
<td>154</td>
<td>1%</td>
<td>8.64</td>
<td>44.68</td>
</tr>
<tr>
<td>Interjection</td>
<td>11</td>
<td>&lt;1%</td>
<td>6.09</td>
<td>34.55</td>
</tr>
<tr>
<td>Wh-determiner</td>
<td>21</td>
<td>&lt;1%</td>
<td>9.05</td>
<td>51.67</td>
</tr>
<tr>
<td>Wh-pronoun</td>
<td>72</td>
<td>1%</td>
<td>7.42</td>
<td>38.79</td>
</tr>
<tr>
<td>Possessive wh-pronoun</td>
<td>0</td>
<td>&lt;1%</td>
<td>--</td>
<td>--</td>
</tr>
</tbody>
</table>

Combinations of Parts of Speech

<table>
<thead>
<tr>
<th>Combinations</th>
<th># Unique Queries</th>
<th>% Unique Queries</th>
<th>Average # Terms</th>
<th>Average # Characters</th>
</tr>
</thead>
<tbody>
<tr>
<td>Noun and Adjective</td>
<td>2,843</td>
<td>24%</td>
<td>4.05</td>
<td>25.15</td>
</tr>
<tr>
<td>Noun and Adverb</td>
<td>414</td>
<td>3%</td>
<td>5.89</td>
<td>32.90</td>
</tr>
<tr>
<td>Noun and Verb</td>
<td>1,564</td>
<td>13%</td>
<td>5.14</td>
<td>29.43</td>
</tr>
<tr>
<td>Noun and Other</td>
<td>2,573</td>
<td>21%</td>
<td>5.17</td>
<td>28.33</td>
</tr>
<tr>
<td>Noun, Adjective and Adverb</td>
<td>145</td>
<td>1%</td>
<td>7.85</td>
<td>45.05</td>
</tr>
<tr>
<td>Noun, Adjective and Verb</td>
<td>397</td>
<td>3%</td>
<td>7.31</td>
<td>42.12</td>
</tr>
<tr>
<td>Noun, Adjective and Other</td>
<td>763</td>
<td>6%</td>
<td>6.55</td>
<td>37.20</td>
</tr>
</tbody>
</table>
I used the results from analyzing query length and part-of-speech usage to manually classify narrow and general queries to develop and evaluate five algorithms that use parts of speech and query length to automatically label queries as either narrow or general. The algorithms were based on the theory that the patterns of part-of-speech usage and query length that on average indicate narrow queries could be used to

<table>
<thead>
<tr>
<th>Grouping</th>
<th># Unique Queries</th>
<th>% Unique Queries</th>
<th>Average # Terms</th>
<th>Average # Characters</th>
</tr>
</thead>
<tbody>
<tr>
<td>Noun, Adverb and Verb</td>
<td>182</td>
<td>2%</td>
<td>8.59</td>
<td>45.24</td>
</tr>
<tr>
<td>Noun, Adverb and Other</td>
<td>190</td>
<td>2%</td>
<td>8.73</td>
<td>45.91</td>
</tr>
<tr>
<td>Noun, Verb and Other</td>
<td>672</td>
<td>6%</td>
<td>7.34</td>
<td>38.96</td>
</tr>
<tr>
<td>Noun, Adjective, Adverb and Verb</td>
<td>63</td>
<td>&lt;1%</td>
<td>12.21</td>
<td>67.75</td>
</tr>
<tr>
<td>Noun, Adjective, Adverb and Other</td>
<td>67</td>
<td>1%</td>
<td>12.04</td>
<td>65.58</td>
</tr>
<tr>
<td>Noun, Adjective, Verb and Other</td>
<td>206</td>
<td>2%</td>
<td>10.02</td>
<td>55.54</td>
</tr>
<tr>
<td>Noun, Adverb, Verb and Other</td>
<td>137</td>
<td>1%</td>
<td>9.96</td>
<td>51.65</td>
</tr>
<tr>
<td>Noun, Adjective, Adverb, Verb and Other</td>
<td>50</td>
<td>&lt;1%</td>
<td>14.06</td>
<td>76.48</td>
</tr>
<tr>
<td>Adjective and Adverb</td>
<td>157</td>
<td>1%</td>
<td>7.47</td>
<td>42.90</td>
</tr>
<tr>
<td>Adjective and Verb</td>
<td>415</td>
<td>3%</td>
<td>7.13</td>
<td>41.08</td>
</tr>
<tr>
<td>Adjective and Other</td>
<td>782</td>
<td>7%</td>
<td>6.47</td>
<td>36.69</td>
</tr>
<tr>
<td>Adjective, Adverb and Verb</td>
<td>67</td>
<td>1%</td>
<td>11.73</td>
<td>65.13</td>
</tr>
<tr>
<td>Adjective, Adverb and Other</td>
<td>70</td>
<td>1%</td>
<td>11.71</td>
<td>63.74</td>
</tr>
<tr>
<td>Adjective, Verb and Other</td>
<td>212</td>
<td>2%</td>
<td>9.86</td>
<td>54.56</td>
</tr>
<tr>
<td>Adjective, Adverb, Verb and Other</td>
<td>52</td>
<td>&lt;1%</td>
<td>13.69</td>
<td>74.46</td>
</tr>
<tr>
<td>Adverb and Verb</td>
<td>198</td>
<td>2%</td>
<td>8.28</td>
<td>43.37</td>
</tr>
<tr>
<td>Adverb and Other</td>
<td>203</td>
<td>2%</td>
<td>8.50</td>
<td>44.49</td>
</tr>
<tr>
<td>Adverb, Verb and Other</td>
<td>146</td>
<td>1%</td>
<td>9.73</td>
<td>50.20</td>
</tr>
<tr>
<td>Verb and Other</td>
<td>705</td>
<td>6%</td>
<td>7.19</td>
<td>39.06</td>
</tr>
</tbody>
</table>
automatically identify narrow queries. These patterns were discovered via analysis of manually labeled queries as presented earlier in this work and were encoded first into algorithms and then into java code. Algorithm 1 uses only query length (Figure 4-1), as represented by number of terms and number of characters in the query.

Algorithm 1

Input:
1) q= query
2) num_terms= number of terms in query
3) num_chars= number of characters in query

Output:
1) Specificity= Narrow, General

Process:
1) If num_terms in q >=4 classify as Narrow
2) Else If num_chars in q >=25 classify as Narrow
3) Else classify as General

Figure 4-1: Algorithm 1 using only query length

Algorithm 2 (Figure 4-2) classifies queries without nouns as general. It classifies queries as narrow if they contain parts of speech for which 84%-99% of queries were manually classified as narrow (as seen in Tables 4-6 and 4-7). The parts of speech included are singular proper noun, Wh-adverb, base form verb, preposition or subordinate conjunction, particle, to and Wh-pronoun.
Algorithm 2

Input:

1) q = query
2) pos[] = list of parts of speech in query
3) Group 1 of Parts of Speech = noun, singular proper; wh-adverb; verb, base form; preposition or subordinate conjunction; particle; to; wh-pronoun

Output:

1) Specificity = Narrow, General

Process:

1) If query has no nouns label it as general
2) Else if the query has a part of speech from Group 1 classify it as narrow
3) Else classify the query as general

Figure 4-2: Algorithm 2, conservative approach using parts of speech with >=84% narrow

Algorithm 3 (Figure 4-3) classifies queries without nouns as general. It classifies queries as narrow if they contain parts of speech for which 75%-99% of queries were manually classified as narrow (as seen in Tables 4-6 and 4-7). The parts of speech included are the broad parts of speech (adverb, verbs, other) and the parts of speech (singular proper noun, comparative adjective, and superlative adjective).
Algorithm 3

Input:

1) q= query
2) pos[ ]= list of parts of speech in query
3) Group 1 of Parts of Speech= adverb; verb; other; noun, singular proper; adjectives, comparative; adjectives, superlative

Output:

1) Specificity= Narrow, General

Process:

1) If query has no nouns label it as general
2) Else if the query has a part of speech from Group 1 classify it as narrow
3) Else classify the query as general

Figure 4-3: Algorithm 3 using parts of speech with >=75% narrow

Algorithm 4 (Figure 4-4) combines the elements of Algorithms 1 and 2 and adds a third element based on both query length and parts of speech. The third element classifies queries as narrow if they contain three or more terms and one of the following parts of speech: the broad parts of speech (adverb or verb) or the parts of speech (singular or mass noun, plural noun, comparative adjectives, superlative adjectives or cardinal numbers).
Algorithm 5 (Figure 4-5) combines the elements of Algorithms 1 and 3 and adds a third element based on both query length and parts of speech. The third element classifies queries as narrow if they contain three or more terms and one of the following broad parts of speech: noun or adjective.
Algorithm 5

Input:

1) q= query
2) num_terms= number of terms in query
3) num_chars= number of characters in query
4) pos[ ]= list of parts of speech query
5) Group 1 of Parts of Speech= adverb; verb; other; noun, singular proper; adjective, comparative; adjective, superlative
6) Group 2 of Parts of Speech = noun, adjective

Output:

1) Specificity= Narrow, General

Process:

1) If query has no nouns label it as general
2) Else if the query has 4 or more terms label it as narrow
3) Else if query has 25 or more characters label it as narrow
4) Else If the query has 3 or more terms and a part of speech from Group 2 classify it as narrow
5) Else If the query has a part of speech from Group 1 classify it as narrow
6) Else classify the query as general

Figure 4-5: Algorithm 5 approach using both query length and parts of speech

I applied all five algorithms to the 25,820 queries in the dataset. 12,011 of the queries were unique. To evaluate the accuracy of the automatic classifications, I compared the classification for 5,115 unique queries to the manual classification from research objective 1 (Table 4-11).
Table 4-11: Summary of results of the five algorithms and their accuracy. The bolded line, Algorithm 4, gives the best result.

<table>
<thead>
<tr>
<th>Algorithm</th>
<th>% of queries classified as narrow</th>
<th>% Unique queries classified as narrow</th>
<th>Accuracy compared to manually labeled subset</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Overall</td>
<td>General</td>
<td>Narrow</td>
</tr>
<tr>
<td>Algorithm 1</td>
<td>28%</td>
<td>32%</td>
<td>74%</td>
</tr>
<tr>
<td>Algorithm 2</td>
<td>15%</td>
<td>16%</td>
<td>56%</td>
</tr>
<tr>
<td>Algorithm 3</td>
<td>28%</td>
<td>31%</td>
<td>66%</td>
</tr>
<tr>
<td><strong>Algorithm 4</strong></td>
<td><strong>50%</strong></td>
<td><strong>55%</strong></td>
<td><strong>80%</strong></td>
</tr>
<tr>
<td>Algorithm 5</td>
<td>53%</td>
<td>58%</td>
<td>77%</td>
</tr>
</tbody>
</table>

The results confirm that query length is an important factor for specificity. By itself, it achieved a 74% accuracy rate. Parts of speech by themselves did not perform as well (as seen in the performance of Algorithms 2 and 3), with accuracy rates of 56% and 66%, respectively. However, combining query length and parts of speech improved the results compared to only using query length. Both Algorithms 4 and 5 outperformed Algorithm 1. I see from comparing the performance of the five algorithms (Table 4-11) that in terms of accuracy there is a trade-off between accurately identifying general queries and accurately identifying narrow queries. In comparing the three top algorithms (Algorithms 1, 4, and 5), I see that the strength of Algorithm 1 is in classifying general queries and its weakness is in identifying narrow queries. Algorithms 4 and 5, on the other hand, are strongest when identifying narrow queries and weakest when identifying general queries.

Overall, I found that I could identify specificity (i.e., narrow or general) automatically and that using multiple factors improves the accuracy of identification. Also, I found that that there is a trade off between classifying general queries and narrow
queries accurately. Both of these findings have implications for future work and search engines, searchers, and advertisers.

**Research Objective 4: Explore whether I can automatically identify distinct user groups by their patterns of part-of-speech-usage**

I first examined how often users use various parts of speech and whether they use multiple different parts of speech (Table 4-12). The data were examined at the query level as opposed to the session level. It was rare for users not to use nouns in their queries. Only 2% of users never used a noun. On the other hand, most users (81%) used multiple parts of speech and not just nouns, with 69% of users using both nouns and adjectives in their queries. Interestingly, adverbs were the least common main group with only 27% of users using them. The categories adjective, verb and other were used by over 50% of users. The least common pattern for users was using all five broad categories of parts of speech, which 21% of users did.

<table>
<thead>
<tr>
<th>Table 4-12: Statistics for Part of Speech Usage By Users</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Grouping</strong></td>
</tr>
<tr>
<td>Used Specific Part of Speech</td>
</tr>
<tr>
<td>Without a Noun</td>
</tr>
<tr>
<td>With Something Other Than a Noun</td>
</tr>
<tr>
<td>With Only Nouns</td>
</tr>
<tr>
<td>Noun</td>
</tr>
<tr>
<td>Noun, singular or mass</td>
</tr>
<tr>
<td>Noun, plural</td>
</tr>
<tr>
<td>Proper Noun, singular</td>
</tr>
<tr>
<td>Proper Noun, plural</td>
</tr>
<tr>
<td>Adjective</td>
</tr>
<tr>
<td>Adjective</td>
</tr>
</tbody>
</table>
Table 4-12: Statistics for Part of Speech Usage By Users

<table>
<thead>
<tr>
<th>Grouping</th>
<th># Users</th>
<th>% Users</th>
</tr>
</thead>
<tbody>
<tr>
<td>Adjective, comparative</td>
<td>29</td>
<td>6%</td>
</tr>
<tr>
<td>Adjective, superlative</td>
<td>64</td>
<td>13%</td>
</tr>
<tr>
<td>Adverb</td>
<td>136</td>
<td>27%</td>
</tr>
<tr>
<td>Adverb</td>
<td>126</td>
<td>25%</td>
</tr>
<tr>
<td>Adverb, comparative</td>
<td>1</td>
<td>&lt;1%</td>
</tr>
<tr>
<td>Adverb, superlative</td>
<td>3</td>
<td>1%</td>
</tr>
<tr>
<td>Wh-Adverbs</td>
<td>43</td>
<td>9%</td>
</tr>
<tr>
<td>Verb</td>
<td>267</td>
<td>53%</td>
</tr>
<tr>
<td>Verb, base from</td>
<td>132</td>
<td>26%</td>
</tr>
<tr>
<td>Verb, past tense</td>
<td>67</td>
<td>13%</td>
</tr>
<tr>
<td>Participle Verb, gerund or present participle</td>
<td>111</td>
<td>22%</td>
</tr>
<tr>
<td>Verb, past participle</td>
<td>65</td>
<td>13%</td>
</tr>
<tr>
<td>Verb, Non-3\textsuperscript{rd} person singular present</td>
<td>162</td>
<td>32%</td>
</tr>
<tr>
<td>Verb, 3\textsuperscript{rd} person singular present</td>
<td>96</td>
<td>19%</td>
</tr>
<tr>
<td>Other</td>
<td>296</td>
<td>59%</td>
</tr>
<tr>
<td>Coordinating conjunctions</td>
<td>74</td>
<td>15%</td>
</tr>
<tr>
<td>Cardinal numbers</td>
<td>160</td>
<td>32%</td>
</tr>
<tr>
<td>Determiners</td>
<td>126</td>
<td>25%</td>
</tr>
<tr>
<td>Existential there</td>
<td>0</td>
<td>&lt;1%</td>
</tr>
<tr>
<td>Foreign words</td>
<td>81</td>
<td>16%</td>
</tr>
<tr>
<td>Preposition or subordinating conjunction</td>
<td>202</td>
<td>40%</td>
</tr>
<tr>
<td>List item marker</td>
<td>20</td>
<td>4%</td>
</tr>
<tr>
<td>Modal</td>
<td>20</td>
<td>4%</td>
</tr>
<tr>
<td>Predeterminer</td>
<td>2</td>
<td>&lt;1%</td>
</tr>
<tr>
<td>Possessive ending</td>
<td>0</td>
<td>&lt;1%</td>
</tr>
<tr>
<td>Personal pronoun</td>
<td>71</td>
<td>14%</td>
</tr>
<tr>
<td>Possessive pronoun</td>
<td>50</td>
<td>10%</td>
</tr>
<tr>
<td>Particle</td>
<td>23</td>
<td>5%</td>
</tr>
<tr>
<td>Symbol</td>
<td>34</td>
<td>7%</td>
</tr>
<tr>
<td>to</td>
<td>59</td>
<td>12%</td>
</tr>
<tr>
<td>Interjection</td>
<td>6</td>
<td>1%</td>
</tr>
<tr>
<td>Wh-determiner</td>
<td>10</td>
<td>2%</td>
</tr>
<tr>
<td>Wh-pronoun</td>
<td>28</td>
<td>6%</td>
</tr>
<tr>
<td>Possessive wh-pronoun</td>
<td>0</td>
<td>&lt;1%</td>
</tr>
</tbody>
</table>

Combinations of Parts of Speech

<table>
<thead>
<tr>
<th>Combination</th>
<th># Users</th>
<th>% Users</th>
</tr>
</thead>
<tbody>
<tr>
<td>Noun and Adjective</td>
<td>344</td>
<td>69%</td>
</tr>
<tr>
<td>Noun and Adverb</td>
<td>136</td>
<td>27%</td>
</tr>
<tr>
<td>Noun and Verb</td>
<td>267</td>
<td>53%</td>
</tr>
<tr>
<td>Noun and Other</td>
<td>295</td>
<td>59%</td>
</tr>
</tbody>
</table>
I desired to detect groups of users based on their part-of-speech usage. To do this, I converted counts of parts-of-speech usage into the more comparable percentage of terms entered by the user that corresponded to each part of speech. I then used K-means clustering to automatically group users into 3, 5, 7, 9, and 15 clusters based on their usage of the five broad categories of parts of speech: noun, adjective, adverb, verb, and other. The best number of clusters was five for two reasons: 1) it minimized the number of clusters that were too small to be analyzed and 2) it produced four clusters that showed distinct patterns for users. Figure 4-6 details the different patterns for the five groups of parts of speech for the first four clusters [Cluster 1: 75 users (15%), Cluster 2: 173 users ...

<table>
<thead>
<tr>
<th>Grouping</th>
<th># Users</th>
<th>% Users</th>
</tr>
</thead>
<tbody>
<tr>
<td>Noun, Adjective and Adverb</td>
<td>127</td>
<td>25%</td>
</tr>
<tr>
<td>Noun, Adjective and Verb</td>
<td>232</td>
<td>46%</td>
</tr>
<tr>
<td>Noun, Adjective and Other</td>
<td>251</td>
<td>50%</td>
</tr>
<tr>
<td>Noun, Adverb and Verb</td>
<td>117</td>
<td>23%</td>
</tr>
<tr>
<td>Noun, Adverb and Other</td>
<td>124</td>
<td>25%</td>
</tr>
<tr>
<td>Noun, Verb and Other</td>
<td>216</td>
<td>43%</td>
</tr>
<tr>
<td>Noun, Adjective, Adverb and Verb</td>
<td>111</td>
<td>22%</td>
</tr>
<tr>
<td>Noun, Adjective, Adverb and Other</td>
<td>118</td>
<td>24%</td>
</tr>
<tr>
<td>Noun, Adjective, Verb and Other</td>
<td>197</td>
<td>39%</td>
</tr>
<tr>
<td>Noun, Adverb, Verb and Other</td>
<td>111</td>
<td>22%</td>
</tr>
<tr>
<td>Noun, Adjective, Adverb, Verb and Other</td>
<td>107</td>
<td>21%</td>
</tr>
<tr>
<td>Adjective and Adverb</td>
<td>127</td>
<td>25%</td>
</tr>
<tr>
<td>Adjective and Verb</td>
<td>232</td>
<td>46%</td>
</tr>
<tr>
<td>Adjective and Other</td>
<td>251</td>
<td>50%</td>
</tr>
<tr>
<td>Adjective, Adverb and Verb</td>
<td>111</td>
<td>22%</td>
</tr>
<tr>
<td>Adjective, Adverb and Other</td>
<td>118</td>
<td>24%</td>
</tr>
<tr>
<td>Adjective, Verb and Other</td>
<td>197</td>
<td>39%</td>
</tr>
<tr>
<td>Adjective, Adverb, Verb and Other</td>
<td>107</td>
<td>21%</td>
</tr>
<tr>
<td>Adverb and Verb</td>
<td>117</td>
<td>23%</td>
</tr>
<tr>
<td>Adverb and Other</td>
<td>124</td>
<td>25%</td>
</tr>
<tr>
<td>Adverb, Verb and Other</td>
<td>111</td>
<td>22%</td>
</tr>
<tr>
<td>Verb and Other</td>
<td>216</td>
<td>43%</td>
</tr>
</tbody>
</table>
(35%), Cluster 3: 189 users (38%) and Cluster 4: 53 users (11%)]. The fifth cluster was too small (10 users) to show a distinct pattern. Table 4-13 gives names and key attributes of the four clusters. Table 4-14 gives the mean values for usage of the five broad categories of parts of speech for each cluster.
1)

![Boxplot Diagram]

2)
Table 4-13: Four cluster of users showing different patterns of parts of speech usage

<table>
<thead>
<tr>
<th>Cluster Name</th>
<th>% of Users</th>
<th>Number of Users</th>
<th>Key Attributes</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cluster 1: All Parts of Speech</td>
<td>15%</td>
<td>75</td>
<td>Primarily use Nouns with the other four categories also being used</td>
</tr>
<tr>
<td>Cluster 2: Mainly Nouns</td>
<td>35%</td>
<td>173</td>
<td>Primarily use Nouns with small numbers of other categories</td>
</tr>
<tr>
<td>Cluster 3: Few Adverbs</td>
<td>38%</td>
<td>189</td>
<td>Primarily use Nouns with some Adjectives, Verbs and Other. Rarely use adverbs</td>
</tr>
<tr>
<td>Cluster 4: Mainly Nouns and Adjectives</td>
<td>11%</td>
<td>53</td>
<td>Primarily use Nouns and Adjectives with smaller numbers of verbs and Other. Rarely use adverbs</td>
</tr>
</tbody>
</table>

Table 4-14: The average usage of the parts of speech for the four clusters of users

<table>
<thead>
<tr>
<th>Cluster</th>
<th>Average Adjective Usage</th>
<th>Average Adverb Usage</th>
<th>Average Noun Usage</th>
<th>Average Verb Usage</th>
<th>Average Other Usage</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cluster 1: All Parts of Speech</td>
<td>0.09</td>
<td>0.02</td>
<td>0.60</td>
<td>0.07</td>
<td>0.22</td>
</tr>
<tr>
<td>Cluster 2: Mainly Nouns</td>
<td>0.02</td>
<td>0.001</td>
<td>0.96</td>
<td>0.01</td>
<td>0.01</td>
</tr>
<tr>
<td>Cluster 3: Few Adverbs</td>
<td>0.27</td>
<td>0.01</td>
<td>0.79</td>
<td>0.05</td>
<td>0.07</td>
</tr>
<tr>
<td>Cluster 4: Mainly Nouns and Adjectives</td>
<td>0.13</td>
<td>0.01</td>
<td>0.66</td>
<td>0.03</td>
<td>0.03</td>
</tr>
</tbody>
</table>
The users in the four clusters show different patterns. Users in Cluster 1 (All Parts of Speech) used ~60% nouns, ~20% other, and smaller numbers of adjectives and verbs with adverbs being the least frequent. In Cluster 2 (Mainly Nouns), users used mostly nouns with a few adjectives. In Cluster 3 (Few Adverbs), users used ~80% nouns, some adjectives, verbs, and others and more rarely adverbs. In Cluster 4 (Mainly Nouns and Adjectives) users used mostly nouns with over 20% adjectives, smaller numbers of verbs and other, and more rarely adverbs. Interestingly, adverb was the most rarely used category for all clusters, and the category noun was the most commonly used. The clusters of users show statistically significant differences between most categories of parts of speech. For the categories noun, verb, and other. The first four clusters are statistically different (Wilcoxon test, p<0.05). For adjective, all but the difference between Cluster 1 and Cluster 3 are statistically different. For adverb, the differences between groups are significant except for that between Cluster 3 and Cluster 4.

Overall, I was successful in automatically separating users into groups that show distinct patterns of part-of-speech usage. This indicates that different users show different patterns of specificity of queries and suggests that for each user or user group there is a continuum of specificity for their queries. Search engines, advertisers and other content providers could leverage this information to treat each user group differently.
Chapter 5

Discussion

Summary of Findings

I have five major findings for this research. First, I confirm that the common association between query specificity and query length is accurate. Longer queries are on average more specific. When automatically classifying queries as narrow or general, I find that query length on its own obtains a 74% accuracy rate. However, this is at the cost of misclassifying many of the narrow queries, indicating that short queries can be narrow and that other factors are needed to fully understand specificity.

Second, I show that the presence of specific parts of speech can be indicative of a query’s specificity. For instance, if a query contains Wh-adverbs there is a 96% chance that it is more specific. However, I find that using only part-of-speech usage for automatically classifying queries as narrow or general is not successful (56% and 66% accuracy). This indicates that while part-of-speech usage is a helpful factor it needs to be combined with other information.

Third, I show that the combination of query length with parts of speech can increase the usefulness of part-of-speech identification for specificity. In automatically classifying queries as either narrow or general, I find that combing query length and parts of speech is better overall than using only one of the two factors. This finding shows that one factor (length) is not sufficient to identify query specificity, but by adding another
factor, parts of speech, one can reach 80% accuracy. This is an encouraging finding as it shows that one may be able to increase the accuracy of classifying queries by adding other factors. Additionally, factors are not necessarily independent of each other. For instance, if 98% of queries have at least one noun, the implication is that, if a query has a part of speech other than a noun, it will almost always have at least two words. Thus, one cannot completely separate the influence of parts of speech from that of query length.

Accordingly, to understand the characteristics that make queries more specific, one has to understand both the individual factors related to specificity of Web search queries and the relationships among factors. The implication of this finding is that the specificity of Web search queries is a multifaceted and complicated question, which necessitates understanding multiple interdependent factors. Future research can take many forms to explore how other factors can be used in combination with parts of speech and query length to improve our understanding of and identification of the specificity of Web search queries. Potential factors to explore include types of named entities in queries, the types of information requested by a query, and the specificity of terms in the query.

Fourth, for classifying query specificity there is a trade-off between correctly classifying queries as narrow and correctly classifying queries as general. Thus the designer of a system would need to decide which type of query is more important to classify correctly. In most cases, classifying the narrow queries correctly seems the best approach, but there may be situations where this is not the case.

Fifth, I found that users show variation in their part-of-speech usage and can be grouped into distinct and meaningful groups based on their patterns of part-of-speech usage. Whereas 52% of queries contained only nouns, only 19% of users use only nouns.
Users do not always use the same grammatical form for every query they enter. Being able to automatically separate users into distinct groups that show different patterns of part-of-speech usage indicates that analyzing the user level is useful as specificity of Web search queries varies among users. Search engines and advertisers could leverage this by building distinct models of specificity for different groups of users. A user could then be placed into a group, and the specificity of their queries evaluated based on the specificity model for that group. Using this information, more relevant content could be provided to users, based on both the specificity of the query and the specificity patterns of their group.

**Limitations of Research**

One limitation of this work is that I only look at two factors. A complete understanding of specificity will require examining the contributions of other factors. A second limitation is that the part-of-speech tagger was designed for longer text documents; because of this, it is not optimal for correctly identifying every part of speech in the shorter texts of queries. Additional work could benefit from the use of a part-of-speech tagger that is specialized for Web queries. One last limitation is that I identified queries as either narrow or general. This is a simplistic definition as there could be many more levels depending on the topic and the context. I understand that specificity is not dichotomous but is instead a continuum, which varies by topic and by user. My approach is but a first step to gaining a more nuanced understanding of this continuum and how the queries’ user and context affect it. Future research could continue exploring how a more
nuanced definition and identification of the specificity of Web search queries could be created.

**Strengths of Research**

There are several strengths of this work. One strength is that I develop a clear method for manually identifying the levels of query specificity that does not depend on the labeler’s subjective view of the queries. A second strength of the work is that I empirically confirm the commonly held view that longer queries are more narrow. Third, this research explores the influence of a second factor, part-of-speech usage both by itself and in conjunction with query length. Fourth, while previous work has relied on manual identification of query specificity, this research developed an automatic method for identifying query specificity. Overall, this research expands the understanding of user’s queries.

**Implications**

Overall, my work has implications for advertisers, marketers, search engines, content providers and users. By understanding the specificity of search queries, I increase the understanding of the user’s underlying goals. Thus one can better fulfill those goals.

There are multiple implications and uses for information on query specificity. First, identifying the specificity of both queries and results would enable search engines or other content providers to match the specificity of a potentially relevant Web page/Ad
to that of a query allowing for more relevant results to be identified and returned. (See Figure 5-1 for how this might work). In this case, results with more general topics and information could be ranked higher for general queries and narrower results higher for narrow queries. Another practical use of identifying query specificity would be associating query specificity with query suggestions allowing narrow queries to receive suggestions that are also narrow and general queries to receive suggestions that were general. Similarly, an increased understanding of query specificity could be used to help users develop queries that clearly show the level of specificity they desire.

Figure 5-1: Examples of possible usage of identification of web search query specificity
Conclusion

With the growing importance of classifying Web search queries and personalization, improving understanding of what types of queries users are entering and what types of information searchers are looking for is critically important. This research explores how one can automatically classify Web search queries as either narrow or general. I first developed a list of nine attributes that I used to create an operationalized definition; however, in this paper, I focused on the relationship between two factors that impact the specificity of Web search queries: parts of speech and query length. I found that particular parts of speech are more common for queries that are narrow and that there is a close relationship between the length of queries, the parts of speech present and the specificity of the query. I used this information to develop and test five algorithms for automatically classifying the queries in the dataset as either narrow or general. I found that users show distinct patterns in parts of speech usage indicating that future work can benefit from creating personalized definitions and continuums of specificity for each user group.

The next step is to examine other factors influencing the specificity of queries, such as the presence of named entities, user intent of queries, types of information requested by queries, and topics. First the factors could be explored individually and then combined with each other and query length and part-of-speech usage. Once the influence of different factors is understood, they can be used to improve the automatic classification of query specificity. Additionally, specificity is a continuum. A more granular classification of specificity is needed to more accurately respond to users’ intent.
Understanding specificity and how it is related to factors such as query length and part-of-speech usage is an important step in automatically identifying the specificity of queries. Lastly, future work can explore practical uses of the identification of query specificity for Web search, implementing the identification in larger programs and using the information to help identify results. One initial step here would be to explore whether search engines treat queries with different levels of specificity differently and whether such searches do a better job of providing results for queries with differing levels of specificity.

Once we determine query specificity, then we improve our understanding of users’ goals and will be better able to fulfill those goals. Better understanding and fulfilling users’ goals will not only benefit users but also search engines, advertisers and all content providers.
References


### Appendix A

#### Dataset Table descriptions

**Table 1: Query Table**

<table>
<thead>
<tr>
<th>Fields Name</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Id</td>
<td>A unique identifier for the record in the table</td>
</tr>
<tr>
<td>Qid</td>
<td>The query id</td>
</tr>
<tr>
<td>user_id</td>
<td>The user id for the user who entered the query. Links the query table to the user table.</td>
</tr>
<tr>
<td>unique_qid</td>
<td>A number that is unique to the query. Links the query table with the unique query table</td>
</tr>
<tr>
<td>Qtime</td>
<td>The time the query was entered</td>
</tr>
<tr>
<td>Query</td>
<td>The actual query</td>
</tr>
</tbody>
</table>

**Table 2: Unique Query Table**

<table>
<thead>
<tr>
<th>id</th>
<th>A unique identifier for the record in the table</th>
</tr>
</thead>
<tbody>
<tr>
<td>unique_qid</td>
<td>A number that is unique to the query. Links the query table and the unique query table</td>
</tr>
<tr>
<td>Query</td>
<td>The actual query</td>
</tr>
<tr>
<td>num_terms</td>
<td>The number of terms in the query</td>
</tr>
<tr>
<td>num_chars</td>
<td>The number of characters in the query</td>
</tr>
<tr>
<td>num_occurs</td>
<td>The number of times the query occurs in the dataset</td>
</tr>
<tr>
<td>num_users</td>
<td>The number of users that use this query</td>
</tr>
<tr>
<td>noun_count</td>
<td>The number of nouns in the query</td>
</tr>
<tr>
<td>nn_count</td>
<td>The number of singular or mass nouns in the query</td>
</tr>
<tr>
<td>nns_count</td>
<td>The number of plural nouns in the query</td>
</tr>
<tr>
<td>np_count</td>
<td>The number of singular proper nouns in the query</td>
</tr>
<tr>
<td>nnps_count</td>
<td>The number of plural proper nouns in the query</td>
</tr>
<tr>
<td>Term</td>
<td>Description</td>
</tr>
<tr>
<td>-----------------</td>
<td>-----------------------------------------------------------------</td>
</tr>
<tr>
<td>adjective_count</td>
<td>The number of adjectives in the query</td>
</tr>
<tr>
<td>jj_count</td>
<td>The number of adjectives in the query</td>
</tr>
<tr>
<td>jjr_count</td>
<td>The number of comparative adjectives in the query</td>
</tr>
<tr>
<td>jjs_count</td>
<td>The number of superlative adjectives in the query</td>
</tr>
<tr>
<td>adverb_count</td>
<td>The number of adverbs in the query</td>
</tr>
<tr>
<td>rb_count</td>
<td>The number of adverbs in the query</td>
</tr>
<tr>
<td>rbr_count</td>
<td>The number of comparative adverbs in the query</td>
</tr>
<tr>
<td>rbs_count</td>
<td>The number of superlative adverbs in the query</td>
</tr>
<tr>
<td>wrb_count</td>
<td>The number wh-adverbs in the query</td>
</tr>
<tr>
<td>verb_count</td>
<td>The number of verbs in the query</td>
</tr>
<tr>
<td>vb_count</td>
<td>The number of base form verbs in the query</td>
</tr>
<tr>
<td>vbd_count</td>
<td>The number of past tense verbs in the query</td>
</tr>
<tr>
<td>vbg_count</td>
<td>The number of gerund/present participle verbs in the query</td>
</tr>
<tr>
<td>vbn_count</td>
<td>The number of past participle verbs in the query</td>
</tr>
<tr>
<td>vbp_count</td>
<td>The number of singular present non-3rd person verbs in the query</td>
</tr>
<tr>
<td>vbz_count</td>
<td>The number of the 3rd person singular verbs in the query</td>
</tr>
<tr>
<td>other_count</td>
<td>The number of other parts of speech in the query</td>
</tr>
<tr>
<td>cc_count</td>
<td>The number of coordinating conjunctions in the query</td>
</tr>
<tr>
<td>cd_count</td>
<td>The number of cardinal numbers in the query</td>
</tr>
<tr>
<td>dt_count</td>
<td>The number of determiners in the query</td>
</tr>
<tr>
<td>ex_count</td>
<td>The number of existential there’s in the query</td>
</tr>
<tr>
<td>fw_count</td>
<td>The number of foreign words in the query</td>
</tr>
<tr>
<td>in_count</td>
<td>The number of the prepositions/subordinating conjunctions in the query</td>
</tr>
<tr>
<td>ls_count</td>
<td>The number of list markers in the query</td>
</tr>
<tr>
<td>md_count</td>
<td>The number of modals in the query</td>
</tr>
<tr>
<td>pdt_count</td>
<td>The number of predeterminers in the query</td>
</tr>
<tr>
<td>pos_count</td>
<td>The number of possessive endings in the query</td>
</tr>
<tr>
<td>Variable</td>
<td>Description</td>
</tr>
<tr>
<td>-------------------</td>
<td>-----------------------------------------------------------------------------</td>
</tr>
<tr>
<td>prp_count</td>
<td>The number of personal pronouns in the query</td>
</tr>
<tr>
<td>prp$_count</td>
<td>The number of possessive pronouns in the query</td>
</tr>
<tr>
<td>rp_count</td>
<td>The number of particles in the query</td>
</tr>
<tr>
<td>sym_count</td>
<td>The number of symbols in the query</td>
</tr>
<tr>
<td>to_count</td>
<td>The number of to’s in the query</td>
</tr>
<tr>
<td>uh_count</td>
<td>The number of interjections in the query</td>
</tr>
<tr>
<td>wdt_count</td>
<td>The number of wh-determiners in the query</td>
</tr>
<tr>
<td>wp_count</td>
<td>The number of wh-pronouns in the query</td>
</tr>
<tr>
<td>wp$_count</td>
<td>The number of possessive wh-pronouns in the query</td>
</tr>
<tr>
<td>specificity_level</td>
<td>Whether the query is narrow or general</td>
</tr>
</tbody>
</table>

**Table 3: User Table**

<table>
<thead>
<tr>
<th>Id</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>user_id</td>
<td>A unique identifier for the record in the table.</td>
</tr>
<tr>
<td>num_queries</td>
<td>The number of queries entered by the user.</td>
</tr>
<tr>
<td>num_unique_queries</td>
<td>The number of unique queries entered by the user.</td>
</tr>
<tr>
<td>num_terms</td>
<td>The number of terms used by the user in queries.</td>
</tr>
<tr>
<td>num_unique_terms</td>
<td>The number of unique terms entered by the user.</td>
</tr>
<tr>
<td>average_num_terms</td>
<td>The average number of terms per a query for the user.</td>
</tr>
<tr>
<td>num_chars</td>
<td>The number of characters used by the user in queries.</td>
</tr>
<tr>
<td>num_unique_chars</td>
<td>The number of unique characters entered by the user.</td>
</tr>
<tr>
<td>average_num_chars</td>
<td>The average number of characters per a query for the user.</td>
</tr>
<tr>
<td>noun_count</td>
<td>The number of nouns in the used by the user.</td>
</tr>
<tr>
<td>nn_count</td>
<td>The number of singular or mass nouns used by the user.</td>
</tr>
<tr>
<td>nns_count</td>
<td>The number of plural nouns used by the user.</td>
</tr>
<tr>
<td>nnp_count</td>
<td>The number of singular proper nouns used</td>
</tr>
<tr>
<td>Variable</td>
<td>Description</td>
</tr>
<tr>
<td>--------------</td>
<td>-----------------------------------------------------------------------------</td>
</tr>
<tr>
<td>nnps_count</td>
<td>The number of plural proper nouns used by the user</td>
</tr>
<tr>
<td>adjective_count</td>
<td>The number of adjectives used by the user</td>
</tr>
<tr>
<td>jj_count</td>
<td>The number of adjectives used by the user</td>
</tr>
<tr>
<td>jjr_count</td>
<td>The number of comparative adjectives used by the user</td>
</tr>
<tr>
<td>jjs_count</td>
<td>The number of superlative adjectives used by the user</td>
</tr>
<tr>
<td>adverb_count</td>
<td>The number of adverbs used by the user</td>
</tr>
<tr>
<td>rb_count</td>
<td>The number of adverbs used by the user</td>
</tr>
<tr>
<td>rbr_count</td>
<td>The number of comparative adverbs used by the user</td>
</tr>
<tr>
<td>rbs_count</td>
<td>The number of superlative adverbs used by the user</td>
</tr>
<tr>
<td>wrb_count</td>
<td>The number of wh-adverbs used by the user</td>
</tr>
<tr>
<td>verb_count</td>
<td>The number of verbs used by the user</td>
</tr>
<tr>
<td>vb_count</td>
<td>The number of base form verbs used by the user</td>
</tr>
<tr>
<td>vbd_count</td>
<td>The number of past tense verbs used by the user</td>
</tr>
<tr>
<td>vbg_count</td>
<td>The number of gerund/present participle verbs used by the user</td>
</tr>
<tr>
<td>vbn_count</td>
<td>The number of past participle verbs used by the user</td>
</tr>
<tr>
<td>vbp_count</td>
<td>The number of singular present non-3rd person verbs used by the user</td>
</tr>
<tr>
<td>vbz_count</td>
<td>The number of the 3rd person singular verbs used by the user</td>
</tr>
<tr>
<td>other_count</td>
<td>The number of other parts of speech used by the user</td>
</tr>
<tr>
<td>cc_count</td>
<td>The number of coordinating conjunctions used by the user</td>
</tr>
<tr>
<td>cd_count</td>
<td>The number of cardinal numbers used by the user</td>
</tr>
<tr>
<td>dt_count</td>
<td>The number of determiners used by the user</td>
</tr>
<tr>
<td>ex_count</td>
<td>The number of existential there’s used by the user</td>
</tr>
<tr>
<td>fw_count</td>
<td>The number of foreign words in the query</td>
</tr>
<tr>
<td>in_count</td>
<td>The number of the prepositions/subordinating conjunctions used by the user</td>
</tr>
<tr>
<td>ls_count</td>
<td>The number of list markers used by the user</td>
</tr>
<tr>
<td>Variable</td>
<td>Description</td>
</tr>
<tr>
<td>------------</td>
<td>-----------------------------------------------------------------------------</td>
</tr>
<tr>
<td>md_count</td>
<td>The number of modals used by the user</td>
</tr>
<tr>
<td>pdt_count</td>
<td>The number of predeterminers used by the user</td>
</tr>
<tr>
<td>pos_count</td>
<td>The number of possessive endings used by the user</td>
</tr>
<tr>
<td>prp_count</td>
<td>The number of personal pronouns used by the user</td>
</tr>
<tr>
<td>prp$ _count</td>
<td>The number of possessive pronouns used by the user</td>
</tr>
<tr>
<td>rp_count</td>
<td>The number of particles used by the user</td>
</tr>
<tr>
<td>sym_count</td>
<td>The number of symbols used by the user</td>
</tr>
<tr>
<td>to_count</td>
<td>The number of to's used by the user</td>
</tr>
<tr>
<td>uh_count</td>
<td>The number of interjections used by the user</td>
</tr>
<tr>
<td>wdt_count</td>
<td>The number of wh-determiners used by the user</td>
</tr>
<tr>
<td>wp_count</td>
<td>The number of wh-pronouns used by the user</td>
</tr>
<tr>
<td>wp$ _count</td>
<td>The number of possessive wh-pronouns used by the user</td>
</tr>
</tbody>
</table>
# Appendix B

## Part of Speech Tags

<table>
<thead>
<tr>
<th>Part of Speech Tag</th>
<th>Description</th>
<th>Example</th>
<th>Broad Label</th>
</tr>
</thead>
<tbody>
<tr>
<td>CC</td>
<td>Coordinating conjunction</td>
<td>and, or</td>
<td>Other</td>
</tr>
<tr>
<td>CD</td>
<td>Cardinal number</td>
<td>1, first</td>
<td>Other</td>
</tr>
<tr>
<td>DT</td>
<td>Determiner</td>
<td>the</td>
<td>Other</td>
</tr>
<tr>
<td>EX</td>
<td>Existential there</td>
<td>there is</td>
<td>Other</td>
</tr>
<tr>
<td>FW</td>
<td>Foreign word</td>
<td></td>
<td>Other</td>
</tr>
<tr>
<td>IN</td>
<td>Preposition/Subordinating conjunction</td>
<td>In, of, like</td>
<td>Other</td>
</tr>
<tr>
<td>JJ</td>
<td>Adjective</td>
<td>white</td>
<td>Adjective</td>
</tr>
<tr>
<td>JJR</td>
<td>Adjective, comparative</td>
<td>whiter</td>
<td>Adjective</td>
</tr>
<tr>
<td>JJS</td>
<td>Adjective, superlative</td>
<td>whitest</td>
<td>Adjective</td>
</tr>
<tr>
<td>LS</td>
<td>List marker</td>
<td>1), 2)</td>
<td>Other</td>
</tr>
<tr>
<td>MD</td>
<td>Modal</td>
<td>could, will</td>
<td>Other</td>
</tr>
<tr>
<td>NN</td>
<td>Noun, singular or mass</td>
<td>chair</td>
<td>Noun</td>
</tr>
<tr>
<td>NNS</td>
<td>Noun, plural</td>
<td>chairs</td>
<td>Noun</td>
</tr>
<tr>
<td>NNP</td>
<td>Proper noun, singular</td>
<td>Jane</td>
<td>Noun</td>
</tr>
<tr>
<td>NNPS</td>
<td>Proper noun, plural</td>
<td>Americans</td>
<td>Noun</td>
</tr>
<tr>
<td>PDT</td>
<td>Predeterminer</td>
<td>both the girls</td>
<td>Other</td>
</tr>
<tr>
<td>POS</td>
<td>Possessive ending</td>
<td>friend's</td>
<td>Other</td>
</tr>
<tr>
<td>PRP</td>
<td>Personal pronoun</td>
<td>I, he, she, it</td>
<td>Other</td>
</tr>
<tr>
<td>PRP$</td>
<td>Possessive pronoun</td>
<td>my, his, her</td>
<td>Other</td>
</tr>
<tr>
<td>RB</td>
<td>Adverb</td>
<td>however, usually, naturally, good, bad</td>
<td>Adverb</td>
</tr>
<tr>
<td>RBR</td>
<td>Adverb, comparative</td>
<td>better</td>
<td>Adverb</td>
</tr>
<tr>
<td>RBS</td>
<td>Adverb, superlative</td>
<td>best</td>
<td>Adverb</td>
</tr>
<tr>
<td>RP</td>
<td>Particle</td>
<td>give up</td>
<td>Other</td>
</tr>
<tr>
<td>SYM</td>
<td>Symbol</td>
<td>&gt;, a, *</td>
<td>Other</td>
</tr>
<tr>
<td>TO</td>
<td>To</td>
<td>to go, to him</td>
<td>Other</td>
</tr>
<tr>
<td>UH</td>
<td>Interjection</td>
<td>uhhuh</td>
<td>Other</td>
</tr>
<tr>
<td>VB</td>
<td>Verb, base form</td>
<td>save</td>
<td>Verb</td>
</tr>
<tr>
<td>VBD</td>
<td>Verb, past tense</td>
<td>saved</td>
<td>Verb</td>
</tr>
<tr>
<td>VBG</td>
<td>Verb, gerund/present participle</td>
<td>saving</td>
<td>Verb</td>
</tr>
<tr>
<td>VBN</td>
<td>Verb, past participle</td>
<td>risen</td>
<td>Verb</td>
</tr>
<tr>
<td>VBP</td>
<td>Verb, singular present</td>
<td>save</td>
<td>Verb</td>
</tr>
<tr>
<td>Tag</td>
<td>Category</td>
<td>Example</td>
<td>Tag</td>
</tr>
<tr>
<td>------</td>
<td>---------------------------------</td>
<td>------------</td>
<td>------</td>
</tr>
<tr>
<td>VBZ</td>
<td>Verb, 3rd person singular present</td>
<td>saves</td>
<td>Verb</td>
</tr>
<tr>
<td>WDT</td>
<td>Wh-determiner</td>
<td>which</td>
<td>Other</td>
</tr>
<tr>
<td>WP</td>
<td>Wh-pronoun</td>
<td>who, what</td>
<td>Other</td>
</tr>
<tr>
<td>WPS</td>
<td>Possessive wh-pronoun</td>
<td>whose</td>
<td>Other</td>
</tr>
<tr>
<td>WRB</td>
<td>Wh-adverb</td>
<td>where, when</td>
<td>Adverb</td>
</tr>
</tbody>
</table>
## Appendix C

### Example queries for the most common parts of speech

<table>
<thead>
<tr>
<th>Part of Speech</th>
<th>Example Query</th>
</tr>
</thead>
<tbody>
<tr>
<td>Coordinating conjunction</td>
<td>“midnight in the garden of good and evil”</td>
</tr>
<tr>
<td>Cardinal number</td>
<td>“411 com”</td>
</tr>
<tr>
<td>Determiner</td>
<td>“giselle the ballet”</td>
</tr>
<tr>
<td>Foreign word</td>
<td>“mi vida loca”</td>
</tr>
<tr>
<td>Preposition/Subordinating conjunction</td>
<td>“the childhood of mark twain”</td>
</tr>
<tr>
<td>Adjective</td>
<td>“free death index”</td>
</tr>
<tr>
<td>Adjective, comparative</td>
<td>“better business bureau”</td>
</tr>
<tr>
<td>Adjective, superlative</td>
<td>“fugee the greatest hits”</td>
</tr>
<tr>
<td>Noun, singular or mass</td>
<td>“tv”</td>
</tr>
<tr>
<td>Noun, plural</td>
<td>“card magnets”</td>
</tr>
<tr>
<td>Proper noun, singular</td>
<td>“chicago yachts”</td>
</tr>
<tr>
<td>Personal pronoun</td>
<td>“how do you make a banner”</td>
</tr>
<tr>
<td>Possessive pronoun</td>
<td>“our lady of lourdes”</td>
</tr>
<tr>
<td>Adverb</td>
<td>“very necessary cd”</td>
</tr>
<tr>
<td>Adverb, comparative</td>
<td>“dresses longer length”</td>
</tr>
<tr>
<td>Adverb, superlative</td>
<td>“texas most wanted”</td>
</tr>
<tr>
<td>Particle</td>
<td>“aol dial up”</td>
</tr>
<tr>
<td>Symbol</td>
<td>“vaughn d j”</td>
</tr>
<tr>
<td>To</td>
<td>“how to cook prime rib”</td>
</tr>
<tr>
<td>Interjection</td>
<td>“hang touch oh oh oh”</td>
</tr>
<tr>
<td>Verb, base form</td>
<td>“ask”</td>
</tr>
<tr>
<td>Verb, past tense</td>
<td>“flower faced bat”</td>
</tr>
<tr>
<td>Verb, gerund/present participle</td>
<td>“dealing with a loved ones will”</td>
</tr>
<tr>
<td>Verb, past participle</td>
<td>“as seen on tv”</td>
</tr>
<tr>
<td>Verb, singular present, non-3rd person</td>
<td>“what kind of job do I want”</td>
</tr>
<tr>
<td>Verb, 3rd person singular present</td>
<td>“what is a hotel”</td>
</tr>
<tr>
<td>Wh-determiner</td>
<td>“which search engines uses meta tags”</td>
</tr>
<tr>
<td>Wh-pronoun</td>
<td>“http means what”</td>
</tr>
<tr>
<td>Wh-adverb</td>
<td>“when was ask jeeves introduced”</td>
</tr>
</tbody>
</table>