Understanding the Specificity of Web Search Queries

Carolyn Theresa Hafernik
College of Information Sciences and Technology
The Pennsylvania State University
University Park, PA 16802 USA
cth132@ist.psu.edu

Bernard J. Jansen
College of Information Sciences and Technology
The Pennsylvania State University
University Park, PA 16802
jjansen@ist.psu.edu

Abstract
Understanding the specificity of Web search queries can help search systems better address the underlying needs of searchers and provide them relevant content. The goal of this work is to automatically determine the specificity of web search queries. Although many factors may impact the specificity of Web search queries, we investigate two factors of specificity in this research, (1) part of speech and (2) query length. We use content analysis and prior research to develop a list of nine attributes to identify query specificity. The attributes are whether a query contains a URL, a location or place name along with additional terms, compares multiple things, contains multiple distinct ideas or topics, a question that has a clear answer, request for directions, instructions or tips, a specific date and additional terms or a name and additional terms. We then apply these attributes to classify 5,115 unique queries as narrow or general. We then analyze the differences between narrow and general queries based on part of speech and query length. Our results indicate that query length and parts-of-speech usage, by themselves, can distinguish narrow and general queries. We discuss the implications of this work for search engines, marketers and users.

Author Keywords
Web Search; Web Search Queries; Query Specificity.

ACM Classification Keywords
H.3.3 [Information Search and Retrieval]

CHI 2013 Extended Abstracts, April 27–May 2, 2013, Paris, France.
ACM 978-1-4503-1952-2/13/04.

Introduction
Differentiating the intent of users can assist information systems in providing users the information they seek.

Examples of narrow and general queries

a) "American Revolution"

b) "Women's Role in the American Revolution"

Query *b* is more narrow and query *a* is more general. Each query is looking for different types of results. As such differentiating between these queries in terms of specificity aids in providing more relevant content to the searcher.

One important aspect of this task is determining the specificity intent of Web search queries. Identifying the specificity of Web search queries entails classifying them on a granularity spectrum of narrow to broad for a given topic.

The goal of this work is to develop better ways to classify queries into general and narrow categories by gaining an understanding of the characteristics of narrow vs. general queries. This research is the first step in being able to automatically identifying query specificity as a spectrum of narrow to general queries.

We define query specificity as how narrow or general the user intent of the query is. For this paper, we define the specificity of queries using nine attributes that we identify as being associated with query specificity. However, other factors can inform the specificity of Web search queries. Here, we investigate two additional factors of specificity, query length and part-of-speech usage.

Related Work

Specificity is an important aspect of search queries [1]; however, it can be difficult to measure. Ultimately, determining query specificity aims to reveal the underlying information needed of a user. Gonzalez-Caro et al. identify specificity as one of 10 aspects of user intent [1]. However, little research has looked explicitly at query specificity. Three areas of research apply to specificity as they provide information about user intent and the detail level of a query.

The first research area focuses on the type of query. Often this research identifies queries that have a narrow focus such as question queries [10] and navigational queries [3]. This is closely related to our research because the groupings analyzed often are one part of the larger whole of narrow queries. Knowledge about differences in query type has implications for specificity. For instance, question queries and navigational queries often have very narrow answers that users are looking for.

The second research area relates to query formulation and reformulation. In query reformulation, there is often a discussion of broad reformulation and narrow reformulations, as in Jansen, Zhang & Spink [2]. Such research requires sequences of queries to identify narrower queries. Our research aims to identify the specificity of queries without needing this sequence of queries.

The last area of research related to specificity is the previous work that deals directly with specificity. Only a few papers, such as [5], deal with identifying the specificity of queries. Research has investigated identifying ambiguous queries and terms [6, 7], which are sometimes linked to general queries. We approach the problem from the opposite direction seeking first to identify characteristics of narrow queries. Specificity is anecdotally associated with length where longer queries are considered more specific [5]; however, there has been surprisingly little empirical investigation of this attribute. For this reason, we chose query length as factor to consider both to confirm the common identification of it with specificity and to serve as a benchmark for testing additional factors.

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?

We first identify attributes of a query that allow us to categorize its' level of specificity. The attributes are detectable by manual analysis of queries. Second, we examine the influence of two additional factors: the identified factor of length and a new factor, part-of-speech usage. Identifying the influence of these additional factors is the first step to being able to automatically identify query specificity.

Research Objectives

Our overall goal is to better understand what specificity means operationally and how it can be detected in queries. We aim to better understand how automatically identifiable factors such as query length and parts of speech relate to specificity. Here we have two objectives:

1. Identify a list of attributes to categorize the specificity of queries and apply these attributes to actual queries.
2. Evaluate the effectiveness of query length and parts of speech in identifying the specificity of queries.

Methodology

Data set

We randomly select 5,115 unique queries from a transaction log containing daily information about search queries from the AOL Search Service from March to May 2006. The transaction log contained 3.5 million searches from 65,000 users. Each record in the transaction log contained information on the anonymous user, date and time a query was submitted, the query, type of search, and the query click URL.

Objective 1: Identify a list of attributes to categorize the specificity of queries and apply these attributes to actual queries.

We analyzed related work on specificity and types of queries. We also examined actual queries from the AOL transaction log using open coding to develop a list of attributes of narrow queries. These attributes were determined by manually examining queries and associating certain attributes with narrow query intents or goals. The nine attributes identified were whether a query contains a URL, a location or place name along with additional terms, compares multiple things, contains multiple distinct ideas or topics, a question that has a clear answer, request for directions, instructions or tips, a specific date and additional terms or a name and additional terms.

Using our 5,115 unique queries, the first author classified the specificity of them as narrow or general based on the characteristics we developed. The nine attributes represent narrow goals for queries and as such queries that had one or more of the attributes, we list above, were labeled as narrow. All other queries were labeled as general. We realize that specificity is not binary. However, we reserve the probabilistic evaluation for future research.

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

Our next step was to investigate narrow and general queries relative to query length and parts of speech. We automatically calculated the number of terms and characters in each query to represent query length.

Attribute	Example Query
URL	''www weather channel com''
Location or place name along with additional terms	''montgomery co tx dental clinics''
Compares multiple things	''us clothing size vs uk size''
Multiple distinct ideas or topics	''prayer saint legal matters''
Question that has a clear answers	''what channel is the draft on''
Request for directions, instructions, tips	''how to check another screen names mailbox''
Specific date and additional terms	''1967 camaro rs''
Number and additional terms	''yamaha sx230''
Name and additional terms	''johnny cash songs''

Table 1. The nine attributes with example queries.

To determine part-of-speech usage, we constructed a Java program that interfaces with an identification tool from Stanford Natural Language Processing Group [8, 9]. The part-of-speech identifier uses the Penn Treebank tag set to label parts of speech [4]. We used these same labels for the subcategories of parts of speech.

We counted the usage of five broad categories of parts of speech: nouns, adjectives, adverbs, verbs and other. We also looked at more narrow parts of speech (e.g. verb, gerund or present participle, coordinating conjunctions etc.). Note that any tagged part of speech that was present in fewer than 25 queries (0.5% of dataset) in the sample was not evaluated due to the small sample size.

Results

First, we detail what makes a query narrow or general and second how narrow and general queries differ in terms of length and part-of-speech usage.

Objective 1: Attributes of Narrow Queries

Definitions of specificity in the literature range from vague [5] to more detailed [1]. The most detailed description of specificity for queries comes from Gonzalez-Caro et al. [1] who divide queries into three groups: specific, medium, and broad with specific queries being those that have a name, data, place, acronym or URL. Medium being those queries that are more general, and broad being queries that contain a very general term.

We analyzed a set of queries from our log to select nine attributes to identify specificity. Because the intent behind the query is important for understanding specificity, we selected attributes that indicate the intent of the user. 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.

The nine attributes (See Table 1 for example queries) were:

1. Query contains a URL
2. Query contains a location or place name along with additional terms
3. Query compares multiple things
4. Query contains multiple distinct ideas or topics
5. Query contains a question that has a clear answer
6. Query contains a request for directions, instructions, tips
7. Query contains a specific date and additional terms
8. Query contains a number and additional terms
9. Query contains a name and additional terms

Our initial classification has two levels: narrow and general. Narrow queries were those that contained one or more of the above attributes. General queries were those that did not contain one of the attributes. We classified a new set of 5,115 queries as narrow or general. Based on this decision tree, we classified 62% (3,103) of the queries as narrow and 38%(2,012) queries as general.

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

Using the classified queries from research objective 1, we analyzed query length and part-of-speech usage

	Narrow	General
# Unique Queries	3,103	2,012
Average # Terms	4.51	2.1
Average # Characters	26.6	13.06

Table 2. Descriptive statistics about general and narrow queries

Broad Part of Speech	% Narrow
Nouns	63%
Adjectives	67%
Adverbs	77%
Verbs	77%
Other	77%

Table 3. Percentages of queries with each broad part of speech that are classified as narrow.

Broad Part of Speech	χ^2
Nouns	$2.9^{N.S.}$
Adjectives	16.3*
Adverbs	37.5*
Verbs	133.5*
Other	166.2*

Table 4. Chi Square for broad parts of speech. * $p<0.0001$, $^{N.S.}$ Not significant.

and whether they differentiate between our previously identified groups of narrow and general queries. On average, narrow queries were twice as long as general queries in terms of the number of terms and the number of characters (unpaired t-test $p<0.0001$) (Table 2).

For parts of speech, we found that for broad parts of speech (Table 3, Table 4) queries containing adverbs, verbs and other categories are more likely to be narrow, whereas queries with nouns and adjectives are closer to the ratio for all queries.

Examining the subcategories of the broad parts of speech (Table 5, Table 6), we see that 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. The exception to this is the category symbols, which is as likely to be general as narrow (Table 5).

Discussion and Implications

Our research has two major findings. First, we confirm that the common association between narrow query specificity and query length is accurate. Longer queries on average have a narrow intent. However, short queries (e.g., the name of a company) also can be narrow indicating that other factors are needed to fully understand specificity. Other factors could include types of named entities in the query or types of words in the query. Secondly, we show that the presence of 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 narrow.

Part of Speech	% Narrow
All Queries	62%
No Nouns	12%
Nouns, Singular or Mass	68%
Nouns, Plural	66%
Nouns, Proper	85%
Adjectives	67%
Adjectives, Comparative	75%
Adjectives, Superlative	76%
Adverb	72%
Wh-Adverbs	96%
Verb, Base form	84%
Verb, Past Tense	77%
Verb, Gerund or Present Participle	72%
Verb, Past Participle	72%
Verb, Non-3rd Person Singular Present	81%
Verb, 3rd Person Singular Present	84%
Coordinating Conjunction	82%
Cardinal Numbers	81%
Determiners	78%
Foreign Words	70%
Preposition or Subordinate Conjunction	86%
Personal Pronoun	76%
Possessive Pronoun	74%
Particple	84%
Symbol	51%
to	91%
Wh-pronoun	99%

Table 5. Percentages of queries with each part of speech that are classified as narrow.

Part of Speech	χ^2
No Nouns	266.9*
Nouns, Singular /Mass	42.6*
Nouns, Plural	13.2*
Nouns, Proper	5.6**
Adjectives	14.2*
Adj., Comparative	$2.6^{N.S.}$
Adj., Superlative	3.9**
Adverb	11.4*
Wh-Adverbs	47.9*
Verb, Base form	62.6*
Verb, Past Tense	14.1*
Verb, Gerund/Present Participle	10.9*
Verb, Past Participle	6.1*
Verb, Non-3rd Person Sing. Present	63.6*
Verb, 3rd Person Sing. Present	52.2*
Coordinating Conj.	26.1*
Cardinal Numbers	58.8*
Determiners	45.9*
Foreign Words	4.7**
Preposition/Sub, Conj.	211.0*
Personal Pronoun	15.1*
Possessive Pronoun	6.4**
Particple	7.5**
Symbol	$2.6^{N.S.}$
to	55.9*
Wh-pronoun	41.0*

Table 6. Chi square values for parts of speech. * p<0.001, **p<0.05, $^{N.S.}$ Not significant.

Our work has implications for advertisers, search engines, content providers, and users. By understanding the query specificity, we increase our understanding of the user's underlying goals. Thus, we can better fulfill those goals by providing more relevant content and suggestions. For instance, associating query specificity with query suggestions allowing narrow queries to receive suggestions that are also narrow. Another use could be for ranking results as in ranking results with more general topics and information higher for general queries and narrower results higher for narrow queries. Advertisers, search engines, and content providers can benefit by being able to target their content to relevant users. Searchers can benefit by receiving content and suggestions more closely tied to their requests and needs.

Conclusions and Future Work

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 parts-of-speech usage is an important step in automatically identifying the specificity of queries. Once we identify query specificity, then we improve our understanding of users' goals and can better fulfill those goals. Advertisers and search engines could use this information to better target ads. Our next step is to apply our findings to develop an algorithm to identify automatically more narrow queries.

References

[1] González-Caro, C., Calderón-Benavides, L., Baeza-Yates, R., Tansini, L., and Dubhashi, D.. Web Queries: the Tip of the Iceberg of the User's Intent. In *Proc. WSDM 2011,* (2011).

[2] Jansen, B.J., Zhang, M., and Spink, A.. Patterns and transitions of query reformulation during Web searching. *International Journal of Web Information Systems 3*, 4 (2007), 328-340.

[3] Lee, W.M. and Sanderson, M.. Analyzing URL queries. *Journal of the American Society for Information Science and Technology 61*, 11 (2010), 2300-2310.

[4] Marcus, M.P., Santorini, B., and Marcinkiewicz, M.A.. Building a Large Annotated Corpus of English: The Penn Treebank. *Computational Linguistics 19*, 2 (1993), 313-330.

[5] Phan, N., Bailey, P., and Wilkinson, R.. Understanding the relationship of information need specificity to search query length. In *Proc. SIGIR* 2007, ACM Press (2007) 709-710.

[6] Sanderson, M., 2008. Ambiguous queries: test collections need more sense. In *Proc. SIGIR 2008*, ACM Press (2008), 499-506.

[7] Song, R., Luo, Z., Nie, J.-Y., Yu, Y., and Hon, H.-W.. Identification of ambiguous queries in web search. *Information Processing & Management 45*, 2 (2009), 216-229.

[8] Toutanova, K., Klen, D., Manning, C.D., and Singer, Y., 2003. Feature-Rich Part-of_Speech Tagging with a Cyclic Dependency Network. In *Proc. HLT-NAACL* 2003, (2003), 252-259.

[9] Toutanova, K. and Manning, C.D., 2000. Enriching the Knowledge Sources Used in Maximum Entropy Part-of-Speech Tagger. In *Proc. EMNLP/VLC 2000,* (2000*)*, 63-70.

[10] White, M.D. and IIivonen, M.. Questions as a factor in Web search strategy. *Information Processing & Management 37*, 5 (2001), 721-740.