HOW PEOPLE SEARCH FOR GOVERNMENTAL INFORMATION ON THE WEB

Bernard J. Jansen
School of Information Sciences and Technology
The Pennsylvania State University

Amanda Spink
School of Information Sciences
University of Pittsburgh

Will Berkheiser
School of Information Sciences and Technology
The Pennsylvania State University

INTRODUCTION

People are now confronted with the task of locating electronic information needed to address the issues of their daily lives. The Web is presently the major information source for many people in the U.S. (Cole, Suman, Schramm, Lunn, & Aquino, 2003), used more than newspapers, magazines, and television as a source of information. Americans are expanding their use of the Web for all sorts of information and commercial purposes (J. Horrigan & Rainie, 2002; J. B. Horrigan, 2004; National Telecommunications and Information Administration, 2002). Searching for information is one of the most popular Web activities, second only to the use of email (Nielsen Media, 1997). However, successfully locating needed information remains a difficult and challenging task (Eastman & Jansen, 2003). Locating relevant information not only affects individuals but also commercial, educational, and governmental organizations.

This is especially true in regards to people interacting with their governmental agencies. Executive Order 13011 (Clinton, 1996) directed the U.S. federal government to move aggressively with strategies to utilize the Internet. Birdsell and Muzzio (1999) present the growing presence of governmental Web sites, classifying them into three general categories, (1) provision of information, (2) delivery of forms, and (3) transactions. In 2004, 29% of American said they visited a government Web site to contact some governmental entity, 18% sent an email and 22% use multiple means (J. B. Horrigan, 2004). It seems clear that the Web is a major conduit for accessing governmental information and maybe services. Search engines are the primary means for people to locate Web sites (Nielsen Media, 1997).

Given the Web's importance, we need to understand how Web search engines perform (Lawrence & Giles, 1998) and how people use and interact with Web search engines to locate governmental information. Examining Web searching for governmental information is an important area of research with the potential to increase our understanding of users of Web-based governmental information, advance our knowledge

of Web searchers' governmental information needs, and positively impact the design of Web search engines and sites that specialize in governmental information.

BACKGROUND

There has been limited large-scale research examining Web searching for governmental information. Croft, Cook, and Wider (1995) present analysis of the use of THOMAS, a governmental system that makes U.S. legislative information available to the public. The researchers report that searchers used very simple queries and appeared to have trouble locating specific bills. The researchers also noted that was some dissatisfaction with the relevance of returned results. Marchionini and Levi (2004) present an on-going study of the Bureau of Labor Statistics, reporting that user interfaces to governmental Web sites require special attention

Hargittai (2003) reports that users look for governmental information in a variety of ways with considerable temporal variance for task completion. The manner in which content is presented often confuses searchers. Hargittai reports that the two major sources of confusion concerned the uniform resource locator and the page design layout. Ceaparu and Shneiderman (2004) investigated alternate ways of organizing governmental statistics, reporting that users were more successful in finding answers when the information was organized into categories rather than an alphabetical listing.

There is a body of research focusing on general Web searching (Jansen, Spink, & Saracevic, 2000; Spink & Jansen, 2004). Spink and Jansen (2004) report that searching for governmental information on the Excite search from 1997 to 2001 was about 1.5% to 3% of all queries. Jansen and Spink (2005) report that searching for governmental information on AlltheWeb.com, a Norwegian-based search engine, was about 2.0% of all queries in a study of datasets from 2001 and 2002. Jansen, Spink, and Pedersen (Jansen, Spink, & Pederson, 2005) report that searching for governmental information on AltaVista in 2002 was about 1.6%. Overall, few Web queries are related to government information; however, as a stand alone category, governmental queries are a sizeable percentage.

Although these studies provide important insights into Web searching, further research is needed that validates these results for the searching of governmental information on Web search engines. This is especially important as Web searching systems are continually undergoing changes and governmental entities are moving more services to the Web (e.g., http://usgovinfo.about.com/library/news/aainternet.htm).

We address this need in the present study by examining a set of queries representing governmental-related information needs to analyze how people are searching for governmental information, including what information they are seeking. We also classify a sub-set of these queries to develop a taxonomy that can asset in the development and organization of governmental Web sites.

FUTURE TRENDS

Research Questions

The research questions driving this study are:

1. What are the characteristics of governmental Web searching?

2. What types of governmental information are people searching for on the Web?

3. How effective are these queries in locating governmental information?

Research Design

Data Collection

To address the first research question, we obtained, and qualitatively analyzed, actual governmental-related queries submitted to the AltaVista Web search engine. For this research question, we are interested in examining the characteristics of the governmental-related queries, investigating areas such as the number of terms in queries, the number of queries in a session, and the use of query operators, among other aspects. Our analyses of searching behavior addressed the following approaches to Web searching behavior.

a) *Query length* - The query length is defined as the length, measured in terms, of the entire search query. This may include Boolean operators.

b) *Session size* – A session is the total amount of queries submitted over a period of time. A session may include only one query, or may extend over a longer period of time and include multiple queries.

c) *Result pages viewed* – Result pages viewed is the number of pages returned by the search engine based on the query submitted that were actually visited by the user.

The third research question involved classifying a sub-set of these queries with a controlled hierarchal vocabulary. This has implications not only for searchers using the Web to locate governmental information but also search engines and Web sites that serve these users.

To investigate our research questions, we gathered data from the AltaVista search engine. In 2002, Alta Vista was the 9th most popular search engine (Sullivan, 2002), had a content collection of 550 million Web pages (Sullivan, 2000), and approximately 5.6 million unique visitors per month. Overall, AltaVista offers a full range of searching options, has an extremely large content collection, and millions of unique visitors per month. After being an independent company for several years, Overture Services purchased AltaVista in 2003 (Morrissey, 2003).

We recorded the queries examined for this study on the AltaVista server on Sunday, 8 September 2002 and span a 24-hour period. We checked news stories from this day to see if any looked as if they may have influence the investigation, namely the term analysis. There did not appear to be a major news stories occurring on this date. However, the date is near the anniversary of the 9-11 attacks.

We recorded the queries in a transaction log that represents a portion of the searches executed on the Web search engine on this particular date. The original general

transaction log contains approximately 1,000,000 records. Each record contains three fields:

1. *Time of Day*: measured in hours, minutes, and seconds from midnight of each day as recorded by the AltaVista server;
2. *User Identification*: an anonymous user code assigned by the AltaVista server
3. *Query Terms*: terms exactly as entered by the given user.

Data Analysis

From the complete transaction logs, we were interested in only those queries that were governmental-related. We therefore culled a subset of queries pertaining to government-related information using a modified snowball sampling technique (e.g., Patton, 1990). More specifically, we started with several seed terms (i.e., government, Congress, taxes, license, etc.) that are central indicators of government-related searching. Using this set of terms, we extracted all records from the transaction log that contained these terms.

We then reviewed the extracted records identifying other terms that frequently appeared. These new terms were then combined with the set of original terms, and from the original transaction log we extracted all records that contained these terms. The process was repeated until the addition of new terms to the set added less than ten new and unique queries. This data set permitted us to address our first research question.

For the third research question, we chose 4,000 queries to classify according to a controlled hierarchal vocabulary. We derived this taxonomy from (Brodie, 2002) and organized a five-level hierarchy, which is:

- Level of Government (c.f., state, local, federal)
- High-Level Request (c.f., forms, information)
- Applicable to (c.f., business, citizen)
- Mid-Level Request (c.f., organization of government, aspect of government, tax service)
- Low-Level Request (c.f., information about government, courts, taxes).

With this controlled vocabulary, we were able to classify the queries and then conducted data analysis on the classified queries in order to find any unique relationships or trends within the classification. A complete list of codes within each category is provided in Appendix A.

Results

Searching Characteristics

We first address research question one, which is: What are the characteristics of governmental Web searching?

Table 1 presents overall descriptive data for the governmental-related queries.

Table 1: Aggregate Results of Searching Characteristics

Sessions	3,021	0.8% of all session			
Queries	4,694	0.4% of all queries			
Terms					
Unique	3,867	1.3% of all unique terms			
Total	18,873	0.6% of all terms			
	Mean	**Max**	**Min**	**Mode**	**SD**
Mean terms per query	4.02	20	1	3	2.06
Terms per query	**Occurrences**	**%**			
1 term	221	5%			
2 terms	802	17%			
3+ terms	3,671	78%			
Total	4,694	100%			
	Mean	**Max**	**Min**	**Mode**	**SD**
Mean queries per user	1.55	16	1	1	1.19
	Occurrences	**%**			
Users modifying queries	938	31%			
Session size	**Occurrences**	**%**			
1 query	2,083	69%			
2 queries	588	19%			
3+ queries	350	12%			
Total	3,021	100%			
	Occurrences	**%**			
Boolean Queries	343	7%			
	Occurrences	**%**			
Terms not repeated in data set	2,384	13%			
	Occurrences	**%**			
Use of 100 most frequently occurring terms	10,337	55%			

The number of governmental sessions was 0.8% of the 369,350 sessions in the entire dataset. The number of queries was 0.4% of the 1,073,388 queries in the data set. The set of governmental queries contained 1.3% of the unique terms in the complete transaction log and represents 0.6% of all terms.

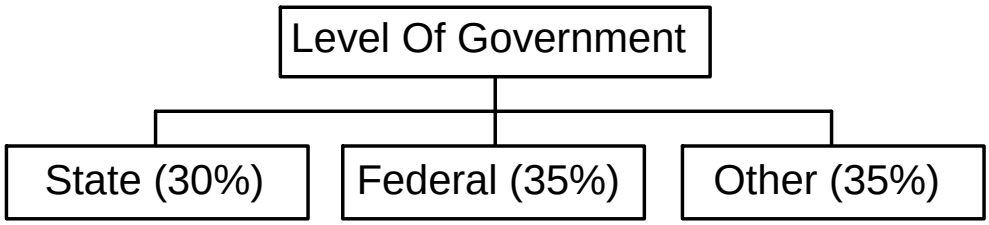
The number of queries in a session and the number of terms in a query are indicators of the complexity of the searcher's information need, with the greater number of queries and terms indicating the greater complexity. The number of terms per query (4.02 terms) is substantially higher than for Web searching in general, which is about 2 terms (Spink & Jansen, 2004). This may indicate that governmental information is more difficult to locate than general Web information or that these searchers have a more refined information need.

The number of one query sessions (69%) is much higher than the entire transaction log, at 47.6% one query sessions(Jansen et al., 2005). The percentage use of Boolean and terms not repeated in the data were comparable to the entire data set. The percentage use of the 100 most frequently occurring terms (55%) was higher than the entire data set (14%), but this may be expected due to the tighter topic domain.

Types of Governmental Information

To address research question two, (i.e, What types of governmental information are people searching for on the Web?), we took 4,000 of the governmental queries and classified them using our five-level hierarchy. The first level of classification is Level of Government. This category includes Federal, State, Local, City, County, etc. Figure 1 presents an overview of the results for this category.

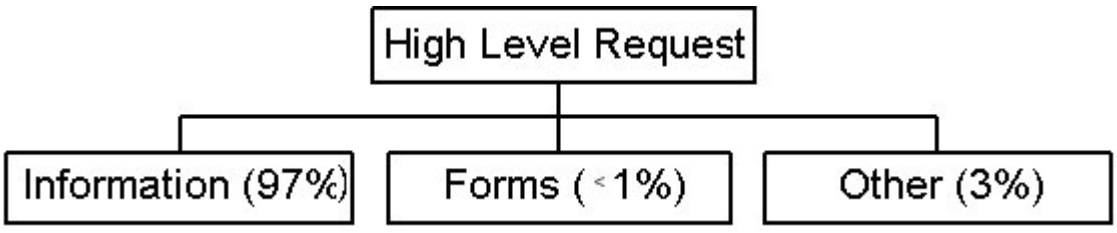
Figure 1. Percentages of queries within level of government categories.



Queries in the Other category were typically searches for local, city or county related governmental information. The other two main levels of government searched for, State and Federal, accounted for 65% of all government-related queries.

The next level of classification is High-Level Request. This level includes requests for Information, Forms, Services, etc. Figure 2 shows the breakdown of the results for this category.

Figure 2. Percentages of queries for type of high level request.



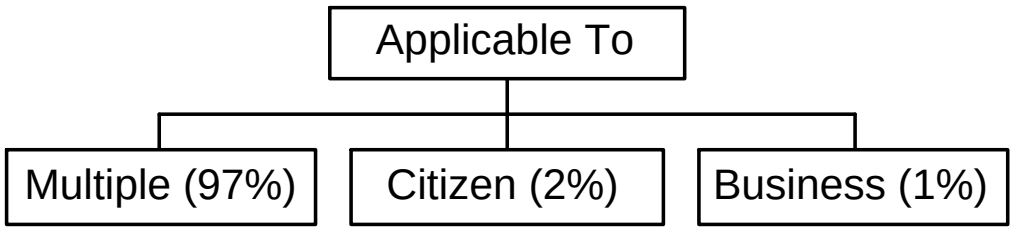
From Figure 2, we see that 97% of the requests were for information, with only 1% for forms. There were very few queries seeking governmental services. This distribution may be because of the preponderance of information of the Web and a lack of awareness of governmental services that are available.

The small percentage of forms at first was surprising, but this may be due to the use of the AltaVista search engine for data analysis. Searchers may be using the search engine

to locate the overall government site (i.e., Internal Revenue Service). Once at the site, they may be searching for the particular form or specific information.

The next category we examined is not necessarily a request level, but an Applicable To level. This level includes Multiple parties, Citizen, Business, etc. Figure 3 shows the breakdown of the results we found for this category.

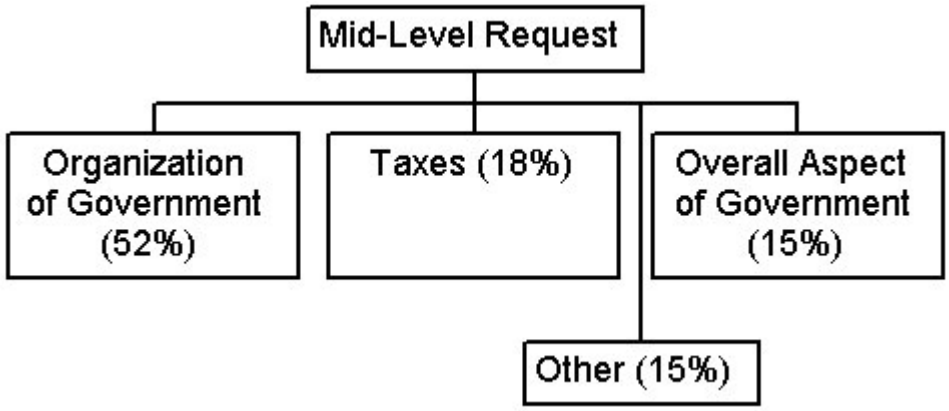
Figure 3. Percentages of queries at applicable to user types.



This classification attempts to address who would be searching for information and could lead to creating an accurate taxonomy or Web site for the user's interest. Unfortunately, when we look at the results of the counts, the largest percentage of the searches is applicable to multiple parties. That is, the query was not specific enough to say who exactly was searching for the information. The other two categories that make up the remainder of the counts are Citizens and Businesses.

We then classified queries at the Mid-Level Request. This level includes requests for information about Organization of Government, Taxes, some Overall Aspect of Government, etc. Figure 4 illustrated the breakdown of the results we found for this category.

Figure 4. Percentages of queries for type of mid level request.

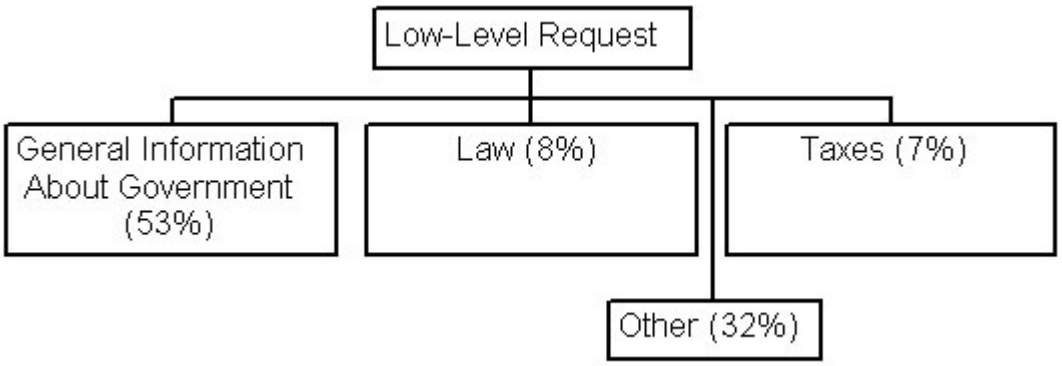


The Mid-Level Request attempts to classify further what exactly is the information need of the query. For example, we can now say that a citizen is looking for basically information about an organization of government. This information is vital when creating accurate taxonomies.

Looking at the percentages in Figure 4, we see that a large percentage of the requests were for an organization of government such as an agency or branch of government. Requests for tax related information is the second largest percentage, followed by a request for some overall aspect of government. Finally, requests for some overall aspect of government was a popular one mostly due to people wondering about general government of the U.S. and policy issues about how government operates.

The fifth level we chose to look at is the Low-Level Request. This level includes General Information about Government, Law, Taxes, etc. Figure 5 shows the breakdown of the results we found for this category.

Figure 5. Percentages of queries for type of low level request.



At this level we come as close as we can to pinpointing what exactly a query was searching for. The percentages again show a strong favoring of General Information about Government, Information about some aspect of Government, etc. This illustrates that mostly what government related queries are after is general information. This is very helpful because when taxonomy is created to organize information.

Other categories that come up as noticeable percentages in the low-level requests are law and taxes. These are both very important aspects of government that deserve to be separated into their own unique categories in order to be further broken down to help direct users to the proper information source.

Effectiveness of Governmental Queries

To address research question three (i.e., How effective are these queries in locating governmental information?), we took the top ten governmental queries and submitted them to Google. We submitted the queries on 26 October 2004 using an automated process. The total submission time was approximately 25 seconds. We downloaded the top 20 results for each of these 10 queries, resulting in 200 results for evaluation. Our criterion for relevance was straightforward, “Was the resulting uniform resource locator (URL) within the gov domain (including US state and mil domains)?” If the URL was in the gov domain, than we considered that result relevant. If the URL was not in the gov domain, than we considered that result no relevant. The results are displayed in Table 2.

Table 2: Aggregate Results of Effectiveness of Governmental Queries

Query	Relevant	%	Non-relevant	%
florida department of revenue	7	35%	13	65%
"White House"	10	50%	10	50%
estate taxes	8	40%	12	60%
+government +"people's party"	1	5%	19	95%
California Department Of Social Services	13	65%	7	35%
North Dakota State Departments "north dakota department"	5	25%	15	75%
clinton, pentagon speech, 1998	3	15%	17	85%
San Bernardino sheriff "sheriff's department"	7	35%	13	65%
ga census +1880	4	20%	16	80%
Washington State Department of Motor Vehicles Seattle"	6	30%	14	70%
	64	32%	136	68%

For our 200 results, 64 (325) were relevant and 136 (38%) were not relevant.

CONCLUSION

From our analysis, it appears that searching for governmental information requires more terms relative to general Web searching. There are also predictable categories of searching for government information on the Web, with most queries applicable to multiple types of searchers. Searchers appear primarily interested in information about government rather than services. It may be difficult to try and isolate particular information objects to sets of users, as queries tend to transcend multiple groups. The relevance of governmental queries is about 32%, with even generous criteria of relevance. This may indicate that governmental agencies have a ways to go to push their information and services to the general population.

Results of this research may lead to better organization of governmental Web information and increase the available of government service y making them easier to locate. This type of research can lead to the creation of more effective government portals.

As governmental agencies continue to push information, forms, and services to their citizenry, it is imperative that they continue to account for the role that search engines have in the process of getting this information to the general population. It is apparent that any categorization will be of limited value to the multitude of users. So, a reliance on indexing and searching will be needed.

REFERENCES

Birdsell, D. S., & Muzzio, D. (1999, February). *Government Programs Involving Citizen Access to Internet Services*. Retrieved 1 June, 2004, from http://www.markle.org/markle_programs/project_archives/2001/dddi.php#report2

Brodie, N. (2002, 4 October). *Taxonomies for Public Access to Government of Canada Information and Services*. Retrieved 19 October, 2004, from http://www.cio-dpi.gc.ca/im-gi/references/class-thes-vocab/taxon/page01_e.asp

Ceaparu, I., & Shneiderman, B. (2004). Finding governmental statistical data on the Web: A study of categorically organized links for the FedStats topics page. *Journal of the American Society for Information Science and Technology, 55*(11), 1008-1015.

Executive Order 13011: Federal Information Technology, (1996).

Cole, J. I., Suman, M., Schramm, P., Lunn, R., & Aquino, J. S. (2003, February). *The UCLA Internet Report Surveying the Digital Future Year Three*. Retrieved 1 February, 2003, from http://www.ccp.ucla.edu/pdf/UCLA-Internet-Report-Year-Three.pdf

Croft, W. B., Cook, R., & Wilder, D. (1995, 11- 13 June). *Providing Government Information on the Internet: Experiences with THOMAS*. Paper presented at the the Digital Libraries Conference, Austin, TX.

Eastman, C. M., & Jansen, B. J. (2003). Coverage, Ranking, and Relevance: A Study of the Impact of Query Operators on Search Engine Results. *ACM Transactions on Information Systems, 21*(4), 383 - 411.

Hargittai, E. (2003). Serving Citizens' Needs: Minimizing Online Hurdles To Accessing Government Information. *IT & Society, 1*(3), 27-41.

Horrigan, J., & Rainie, L. (2002, 29 December). *Counting on the Internet: Most find the information they seek, expect*. Retrieved 24 June, 2004, from http://www.pewinternet.org/PPF/r/80/report_display.asp

Horrigan, J. B. (2004, 24 May). *How Americans Get in Touch With Government*. Retrieved 14 Spetember, 2004, from http://www.pewinternet.org/pdfs/PIP_E-Gov_Report_0504.pdf

Jansen, B. J., & Spink, A. (2005). An Analysis of Web Searching By European Alltheweb.com Users. *Information Processing and Management, 41*(2), 361-381.

Jansen, B. J., Spink, A., & Pederson, J. (2005). Trend Analysis of AltaVista Web Searching. *Journal of the American Society for Information Science and Technology, 56*(6), 559-570.

Jansen, B. J., Spink, A., & Saracevic, T. (2000). Real Life, Real Users, and Real Needs: A Study and Analysis of User Queries on the Web. *Information Processing and Management, 36*(2), 207-227.

Lawrence, S., & Giles, C. L. (1998). Searching the World Wide Web. *Science, 280*(3), 98-100.

Marchionini, G., & Levi, M. (2004). Digital Government Information Services: The Bureau of Labor Statistics Case. *Interactions: New visions of human-computer interaction, 10*(4), 18-27.

Morrissey, B. (2003, February 18, 2003). *Overture to Buy AltaVista*. Retrieved 16 May, 2003, from http://www.internetnews.com/IAR/article.php/1587171

National Telecommunications and Information Administration. (2002). *A Nation Online: How Americans are Expanding their Use of the Internet*. Washington, D.C.: U.S. Department of Commerce.

Nielsen Media. (1997). *Search Engines Most Popular Method of Surfing the Web*. Retrieved 30 August, 2000, from http://www.commerce.net/news/press/0416.html

Patton, M. Q. (1990). *Qualitative Evaluation and Research Methods* (Second ed.). Newbury Park, CA: Sage Publications, Inc.

Spink, A., & Jansen, B. J. (2004). *Web Search: Public Searching of the Web*. New York: Kluwer.

Sullivan, D. (2000). *Search Engine Sizes*. Retrieved 30 August, 2000, from http://searchenginewatch.com/reports/sizes.html

Sullivan, D. (2002, 23 Feb). *Nielsen / NetRatings Search Engine Ratings*. Retrieved 6 January, 2002, from http://www.searchenginewatch.com/reports/netratings.html

Key terms: Web searching, e government, online governmental information, searching for governmental information

Appendix A

Table 3: Level of Government

Code	Description
City	request to a city government
Federal	request to Federal Government
INT	request for information on non-US government
Local	request to a local government
Multiple	request that can be from multiple levels
NA	can't tell from information
School	request to a school board
State	request to a State Government

Table 4: High-Level Request

Code	Description
Forms	a request for forms
Information	a request for basically information
Multiple	a request for multiple things
Services	a request for services

Table 5: Applicable to

Code	Description
Business	a commerce request
Citizen	a request for information of interest to a citizen
CT	can't tell
Multiple	a request that may be from multiple stakeholders
Non-citizen	a request of interest to a non-citizen
Other	

Table 6: Mid-Level Request

Code	Description
art	information on some overall aspect of government
bus_with	non-citizen - doing business with US
financing	business
health	citizen
hr	business human resource
job	citizen
law	multiple
management	multiple
Multiple	request for multiple things
org	looking for a organization of government
policy	multiple
regulations	business

Table 6: Mid-Level Request

Code	**Description**
seniors	citizen
start-up	business start-up
taxes	US citizen
tech	multiple - technology
to_US	non-citizen - coming to US
world	non-citizen - coming to US
youth	citizen

Table 7: Low-Level Request

Code	**Description**
forms	specific type of form
stats	general stats about US
info	general information about government
policy	specific policy
courts	general information
edu	Education