Understanding domain "relevance" in web search

Shanu Sushmita, Hideo Joho, Mounia Lalmas, and Joemon M. Jose
Department of Computing Science, University of Glasgow
Sir Alwyn Williams Building, Lilybank Gardens, Glasgow, G12 8QQ, UK.
{shanu,hideo,mounia,jj}@dcs.gla.ac.uk

ABSTRACT

This paper presents results of a study that analyses people's behavior in accessing different domains (e.g., images, movies, blogs, etc.) in web search. Our results show that the information seeking process of web users involves accessing information from various domains, which suggests the need to provide results from various domains in an aggregated manner. This study also indicates the existence of associations between query categories and domains, which suggests that it is important (and possible) to select domains that are "relevant" to the query.

Categories and Subject Descriptors

H.3.3 [**Information Search and Retrieval**]: [Search process]

General Terms

Experimentation, Human Factors

Keywords

information domain, domain relevance, domain combination, query category, aggregated search

1. INTRODUCTION

Studies reported in [5] show that 80% of web queries are non-navigational; users are usually seeking general information on broad topics, e.g. "global warming", "nutrition". There is typically not a single web page that contains all the information sought; indeed, users with non-navigational queries typically try to assimilate information from multiple web pages, now increasingly from different sources.

Meanwhile, the diversity and complexity of contents available on the web have dramatically increased in recent years. Multimedia content such as images, videos, maps, voice recordings has been published more often than before. Document genres have also been diversified, for instance, news, blogs, FAQs, wiki.

Such growth in the diversity of information on the web suggests investigating two important research questions. Firstly, given such growth in information diversity, do users actually access information from various sources to satisfy their information need? Secondly, how does the presentation of information sources influence the information seeking behavior of users?

WWW2009, April 20-24, 2009, Madrid, Spain.

These diversified information sources are often dealt with in a separated way in search results. In general, users have to switch search domains to access different sources. Recently, there has been a growing interest in finding effective ways to aggregate these information sources in a unified fashion. So-called aggregated search investigated by the like of Yahoo! (Alpha Yahoo!) and Google (Universal Search) are providing search results from several sources in a single result page, where, for example, the results from each source are shown within a panel dedicated to that source. Such an example is shown in Figure 1.

Figure 1: Aggregated search of Yahoo! Alpha presenting results from different domains

The aim of this paper is to address the first research question, where evidences of users accessing information from various sources have been collected. We also discuss briefly the second research question, where a possible way of aggregating results based on users' query is suggested.

This paper presents results of a study that analyses people's behavior in accessing different information sources, called domains in this paper. The objectives of this analysis is to understand users' needs on accessing multiple domains, to find association between domains that are accessed together in search sessions, and to gain an insight into the properties of queries that can characterise the relevant domains. The study makes use of log data where all results, whatever the information source, were shown within a single list of results. This allows us to investigate the "relevance" of domains without the effect (eventually) coming from an aggregated search result presentation, such as shown in Figure 1.

We consider this study as the first step towards the devel-

opment of aggregating results from domains "relevant" to a given query.

2. METHODOLOGY

The following section describes the methodology we followed using the Microsoft 2006 RFP Dataset, a query log of 15 million queries from US users sampled over one month.

2.1 Data set

We hypothesized that search result aggregation was most useful for supporting non-navigational queries [4]. To focus on the non-navigational queries in the dataset, we separated the data into two sets based on the number of clicks made within a single session. More specifically, we made a single-click set, which had only one click in a session, and multiple-click set, which had more than one click in a session. Although this was a very simple method to separate navigational queries from the others, single clicks are one of the main properties of the navigational queries [6]. As a result, we had 3,218,588 single-click sessions (27%) and 8,932,479 multiple-click sessions (73%). Our analysis mainly focused on the multiple-click set. We did not include zero-click queries in our analysis.

2.2 URL analysis

We used the following set of pattern rules to identify the domains of click-through documents. For example, if a click-through URL contained a string *movies*, it was assumed that the main content of the clicked document was a movie.

Image: `/img/ /images/ /image/ /pictures/ /picture/ /photo/ /photos/`

Video: `/vid/ /video/ /videos/ /movie/ /movies/`

Wikipedia: `/wiki/`

News: `/news/`

Blogs: `/blogs/ /blog/`

Audio: `/audio/ /audios/`

Map: `/map/ /maps/`

Web + Others: URLs that did not match any of the above

While the patterns may not be exhaustive to identify all pages belonging to a domain, we considered this as a reasonable approximation of the distribution of different domains in the dataset. The generation of a more comprehensive set of patterns is the aim of our future work. We will also look at more domains in our next study.

3. RESULTS AND ANALYSIS

This section presents the results of our analysis.

3.1 Distribution of information domains

First, we counted the number of clicked URLs that matched with any of the domains in the multiple-click set. There were 274755 click-through URLs (3%) that matched with one of the seven domains (image, video, wikipedia, news, blogs, audio, and map) and the rest was classified as Web + Others. Figure 2 shows the distribution of the seven domains based on the matched URLs.

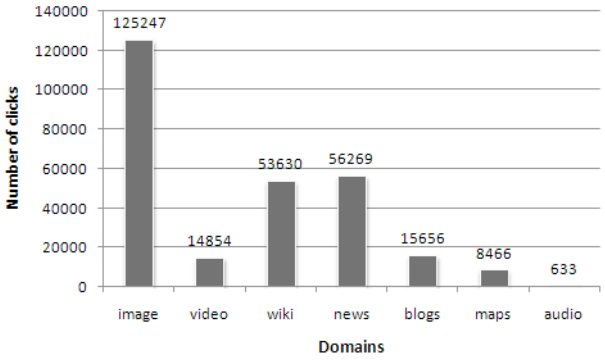

Figure 2: Distribution of domains.

As we can see, images were the most frequent domain followed by wikis and news (similar findings were reported in [1]).

3.2 Association of information domains

We were also interested in the co-occurrence of information domains within sessions since this would indicate the needs for identifying domain intent in aggregated search results. For each session, we recorded the presence of domains in a binary form. Table 1 shows the frequency of co-occurred information domains.

Aggr.	Freq	N	V	B	M	I	Wi	A	W+O
2	42990	1	0	0	0	0	0	0	1
2	41353	0	0	0	0	1	0	0	1
2	40328	0	0	0	0	0	1	0	1
2	11852	0	0	1	0	0	0	0	1
3	1233	1	0	0	0	0	1	0	1
3	1036	0	0	0	0	1	1	0	1
3	829	1	0	1	0	0	0	0	1
3	654	1	0	0	0	1	0	0	1
3	495	1	1	0	0	0	0	0	1
4	61	1	0	0	0	1	1	0	1
4	49	1	0	1	0	0	1	0	1
4	34	1	1	0	0	0	1	0	1
4	20	1	1	1	0	0	0	0	1

Aggr.: # of domains; N: News; V: Video; M: Maps
I: Image; Wi: Wiki; A: Audio; W+O: Web+Others

Table 1: Co-occurrence of domains.

We can make several observations. First, the domains such as News, Images, and Wiki pages equally co-occurred with the web pages (the top three of Aggr. = 2), although their distribution of clicks varied (see Figure 2). Firstly, this suggests that there could be a high level of need for aggregating these domains with web pages in search results. The rest of the domains followed a similar pattern to the distribution figure.

Second, for the co-occurrence of three domains, the News domain seems to be a hub between the web pages and other domains. Third, the frequency of co-occurrence of four domains was very low, suggesting that searchers might not prefer to see too many domains aggregated in search results. However, this observation could be significantly influenced by how results from multiple domains were presented, as well as the quality of these results and domains (e.g. relevance).

3.3 Query category and domain selection

We also collected the 100 most frequent unique queries for 12 different domain combinations. Combinations of 4, 3 and 2 domains were considered (as shown in table 2). Each query from the selected domain combination was then manually assigned to one of the categories defined in the open directory project (OPD) [2]. Any query that did not fit into an existing category was assigned to a 'others' category. The categorization of queries was done to identify associations between query categories and domains, including combinations of domains.

Four	Three	Two
news-blog-wiki-web	image-wiki-web	image-web
news-video-image-web	news-blog-web	news-web
news-video-wiki-web	news-wiki-web	video-web
news-image-wiki-web	news-video-web	blog-web

Table 2: Domain combinations

Two observations can be made. Firstly, there are evidences of associations between the category of a query and the domains. For example, the number of clicks for all combinations of 'news-blog-web' increases for the 'business' category. Similar observation was made for the category 'science' where, combinations with 'image' were preferred. However, associations for categories such as recreational, news, reference and shopping were not clear. For example, the 'recreational' category maintained a high score for all combinations, although some preference for the 'video' domain was observed for this category. The percentage of clicks per category with respect to domain combinations is shown in Table 3.

We also looked at the number of domains with respect to the query category. Results for some categories were very distinct, whereas for others, they were quite mixed (see Table 4). For example, for the category 'business' the highest number of clicks (60%) were from the combination of *three* domains, this number was less for *four* (28.89%) and *two* (11.11%) domain combinations. This indicates that most users searching information related to business gathered information from more than one domain. Similarly, the number of domains was high for the category 'health' , where most people viewed results from more than *three* domains.

Our observations from this analysis give an indication that there are associations between a query category and domains, in terms of which domains and the number of domains. Although our analysis does not yet provide clear conclusions, it provides some insight into how the information seeking behavior of web users in terms of the domains accessed is related to the query category.

3.4 Effect of rank positions

We also looked at the rank positions of clicked domains. While there is a general trend of higher ranked documents being clicked more frequently than lower ranked documents, we have limited understanding of the effect of domains on this behaviour. Figure 3 shows the distribution of ranked positions (one to five) within each domain. Related findings were obtained in a study [1]

As one can see, most domains follow the general click-through pattern, which is a high frequency of the top 1 with monotonically decreasing frequency. This suggests that the selection of domains may strongly be affected by the ranking positions. An exception was for the image domain. This domain had a similar level of click-through rate across the rank positions, although it still follows the general decreasing trend. This suggests that the influence of ranking positions on the selection of the image domain was weaker than with the other domains.

Note that for this analysis we used an entire log set which combined the single-click and multiple-click sets.

Category	Four	Three	Two
arts	32	52	16
business	28.89	60	11.11
computers	0	0	100
health	57.78	28.89	13.33
home	0	50	50
news	23.02	52.38	24.6
recreation	41.7	26.14	32.16
reference	44.83	41.38	13.79
world	25.37	59.7	14.93
science	15.22	36.96	47.83
shopping	29.41	28.24	42.35
society	10.61	16.67	72.73
sports	11.11	18.52	70.37
adult	26.47	22.06	51.47
others	60.61	9.09	30.3

Table 4: Percentage of query category with respect to the four, three and two domain combinations (highlighted values show strong evidences of 'number of domains' accessed w.r.t to a category)

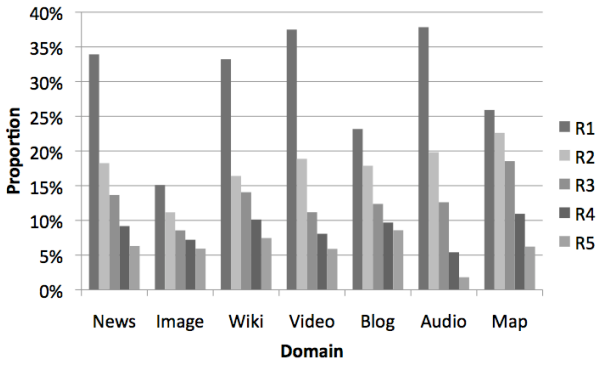

Figure 3: Effect of rank positions (Rank 1-5).

3.5 Effect of click orders

Further, we analysed the effect of click orders on the access patterns of different domains. The objective was to examine whether a certain domain tended to be accessed earlier or later in search sessions. For this analysis, we counted the number of clicks and recorded the click order for those sessions that had at least three click-through documents. 6788 sessions had more than three clicks. The result of our analysis is shown in Figure 4.

As can be seen, most domains had a similar level of frequency of clicks across the click order in sessions. There was a trend of decreasing frequency as the click number increased. Image domain, however, had a very different pat-

	NBWiW	NVIW	NVWiW	NIWiW	IWiW	NBW	NWiW	NVW	IW	NW	VW	BW
arts	8	0	4	20	12	0	40	0	16	0	0	0
business	27.78	0	0	1.11	2.22	38.89	7.78	11.11	0	8.89	2.22	0
computers	0	0	0	0	0	0	0	0	77.78	0	0	22.22
health	8.89	0	0	48.89	15.56	0	13.33	0	0	8.89	0	4.44
home	0	0	0	0	50	0	0	0	50	0	0	0
news	5.56	1.59	8.73	7.14	0.79	22.22	16.67	12.7	0	23.02	0.79	0.79
recreation	7.47	19.29	8.71	6.22	4.77	4.36	6.22	10.7	8.71	6.02	13.97	3.53
reference	13.79	6.9	13.79	10.34	6.9	10.34	10.34	13.79	3.45	0	10.34	0
world	17.91	0	4.48	2.99	46.27	0	13.43	0	0	11.94	1.49	1.49
science	0	0	0	15.22	26.09	6.52	0	4.35	41.3	0	0	6.52
shopping	8.24	0	1.18	20	8.24	3.53	11.76	4.71	11.76	17.65	0	12.94
society	0	1.52	3.03	6.06	7.58	1.52	4.55	3.03	10.61	3.03	6.06	53.03
sports	3.7	0	7.41	0	14.81	0	0	3.7	0	3.7	22.22	44.44
adult	1.47	1.47	23.53	0	2.94	8.82	0	10.29	7.35	2.94	23.53	17.65
others	3.03	3.03	54.55	0	0	0	3.03	6.06	12.12	6.06	0	12.12

Table 3: Percentage clicks for domain-query category, where N= news, B= blog, Wi= Wiki, I=image, V= video, W= web+others (highlighted values show strong evidences for 'category and domain combination')

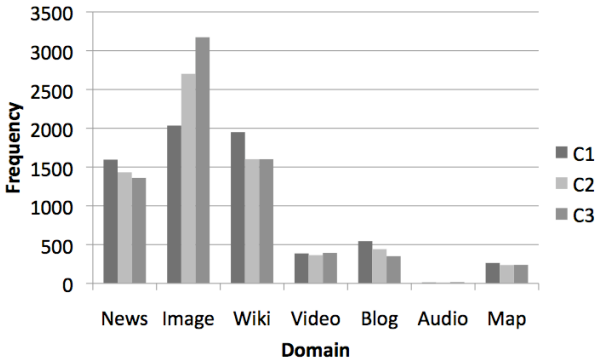

Figure 4: Effect of click order (Click 1-3).

tern. People seem to click the image domain more frequently as the search session progresses. This suggests that the value of the image domain can increase as the search process progresses.

4. CONCLUSION

The aims of this research were to investigate whether there is a need to provide aggregated search, and whether there are associations between a query and the domains so that the most "relevant" domains can be identified for a given query. With respect to the first aim, we can see that, although by large standard web pages are being accessed (clicked), the percentage of non-standard web pages that are clicked (news, blogs, images, etc) is not negligible. It should be noted that, in the log data used in our study, all results (for a given query), whatever their domain, were shown within a single ranked list of results. This indicates that returning non-standard web results is of benefit to the user.

Whether all result types should be merged into a single list is not clear. Note that in this study, we did not differentiate wether results from domains, e.g images were returned as html pages in the ranked list, or obtained explicitly using the respective domain search. Also, whether it is because all result types were returned within a single list that some non-standard web results were clicked is also not clear. The aggregated search paradigm followed by some search engines do not mix the types of results. In a user study, the effectiveness of an aggregated result page, where results from a domain were shown in a dedicated panel to that domain, was compared to a plain ranked list. It was observed that producing such an aggregated result page in response to a non-navigational query improved information access and made task completion quicker. Also, this aggregated search paradigm was often preferred over the plain ranked list result [3].

With respect to our second aim, we observed that there is an association between a user query (category) and domains. Also, for certain query category, there is a dominance of specific domains, which indicates that not all domains are "relevant" to all information needs (here approximated as the queries submitted within a session). The aim of our future research will be to automatically identify the most suitable combinations of domains for any given query.

Acknowledgements

This work was carried out in the context of research partly funded by a Yahoo! Research Alliance Gift.

5. REFERENCES

[1] http://www.iprospect.com/premiumPDFs/researchstudy_apr2008_blendedsearchresults.pdf.
[2] http://dmoz.org.
[3] Under blind review.
[4] A. Broder. A taxonomy of web search. *SIGIR Forum*, 36(2):3–10, 2002.
[5] B. J. Jansen, D. L. Booth, and A. Spink. Determining the informational, navigational, and transactional intent of web queries. *Inf. Process. Manage.*, 44(3), 2008.
[6] D. E. Rose and D. Levinson. Understanding user goals in web search. In *Proceedings of the 13th international conference on World Wide Web*, 2004.