Dynamics of Genre and Domain Intents

Shanu Sushmita, Benjamin Piwowarski, and Mounia Lalmas

University of Glasgow
{shanu,bpiwowar,mounia}@dcs.gla.ac.uk

Abstract. As the type of content available on the web is becoming increasingly diverse, a particular challenge is to properly determine the types of documents sought by a user, that is the domain intent (e.g. image, video) and/or the genre intent (e.g. blog, wikipedia). In this paper, we analysed the Microsoft 2006 RFP click dataset to obtain an understanding of domain and genre intents and their dynamics i.e. how intents evolve within search sessions and their effect on query reformulation.

1 Introduction

The diversity of the content available on the web has dramatically increased in recent years. Multimedia content such as images, videos, maps, has been published more often than before. Document genres have also been diversified, for instance, news, blogs, FAQs, wikipedia. Such growth in the diversity of information on the web raises two main questions. First, do users actually access various categories of documents to satisfy their information need? Second, are there particular patterns in how users access these various categories of documents?

Understanding the information need behind a user query, i.e. the *query intent*, is an important goal in web search. There are different ways to classify query intents. A query intent may refer to the type of interaction, e.g. navigational, transactional or informational. It may also relate to some predefined general topics, e.g. travel, sport, shopping. Finally, it may refer to the category of information being sought, e.g. image, video, blog. In this paper, we are concerned with the latter, more precisely, in *domain* (e.g., image, video) and *genre* (e.g., blog, wikipedia) intents.

In this paper, we analysed the Microsoft 2006 RFP click dataset to understand query intents in terms of domain and genre. We looked at three domains, namely, image, video, and map, and three genres, namely, news, blog and wikipedia. All other categories of intent were viewed as standard "web", i.e. the typical web search result. These six "non-web"[1] categories of intent were chosen on the basis of a survey [4], which shows that images, news, and videos were the three most frequently accessed "non-web" results. Map and wikipedia were chosen because results of these categories are now frequently included within

[1] We use the term "non-web" to distinguish these documents to the standard web documents. In current search engine terminology, these "non-web" documents could be retrieved directly from verticals.

the top ten result list by major search engines. This paper has two parts. We first propose a methodology to identify domain and genre intents, allowing us to learn whether users actually access documents of various categories to satisfy their information need. We then study the dynamics of these intents to uncover patterns in how users access documents of various categories.

2 Data Set

We used the Microsoft 2006 RFP Dataset, containing 13,279,076 log entries [6] corresponding to a subset of web search logs from US users spanning over a month. Each log entry corresponds to a click and contains the following information:

1. The **timestamp** (time and date), used to order the clicks within a session.
2. A unique **session id** representing a search session. A session is the entire series of queries, one to several, submitted to the search engine by a user over some given time.
3. A unique **query id** given to each distinct query submitted during a search session.
4. The **query terms** used by the search engine to display results.
5. The **URL** of the clicked result.

We used three sources of evidence: (1) the query itself, more precisely the use of *intent-specific terms* such as "video", "map", etc; (2) the clicked URLs; and (3) the titles of the corresponding web documents. Previous work, e.g. [3], has shown that result snippets (title, URL and excerpt) of the clicked web documents could be used to determine query intents. For instance, a log entry in our dataset corresponds to a document with title "images and diagrams of human muscular system" after having entered the query "muscular system". The term image in the document title, and the fact that the user has accessed that document, may be a good indication that the user is looking for images.

Total number of entries	6,637,590
Total number of sessions	3,960,541
No of sessions with 1 click	2,654,794
No of sessions with 2 clicks	721,223
No of sessions with 3 clicks	282,980
No of sessions with 4 clicks	132,834
No of sessions with 5 clicks	68,403
No of sessions with > 5 clicks	100,307
Average duration of a session	00:03:44

Table 1. Statistics about the click dataset used in our study.

Number of distinct intents	**Original**	**Random**
Seven	0.000	0.000
Six	0.000	0.000
Five	0.000	0.002
Four	0.005	0.034
Three	0.121	0.470
Two	3.385	6.354
One	96.489	93.140

Table 2. Percentage of sessions with one, two, ... , six and seven different intents.

We did not use document excerpts as we did not have access to them. As the data set did not contain the titles of the clicked documents, we had to

retrieve these. We therefore downloaded the clicked documents. Given the time lag between when the dataset was created and our download of the clicked documents (approximately 3 years), we were able to obtain the titles for only 50% of the clicked documents. We then used those sessions for which we could download the titles for all clicked documents. The statistics reported in Table 1 correspond to these log entries.

3 Determining domain and genre intents

The first stage of our work is to determine the intended domains and genres for given queries. We first used a rule-based classifier, whose output was used to build the features used by a machine learning classifier.

Intent	Rule	SVM	Both	N	image	video	blog	map	wiki	news	web
image	1.10	0.32	1.42	98	**0.30**	**0.18**	0.01	0.00	0.00	0.01	**0.49**
video	0.64	0.86	1.50	131	0.02	**0.46**	0.02	0.00	0.00	0.03	**0.47**
news	0.71	1.43	2.14	62	**0.08**	**0.12**	**0.27**	0.00	0.01	0.03	**0.49**
map	1.50	0.03	1.54	25	0.00	0.00	0.00	**0.63**	0.00	0.00	**0.37**
blog	0.03	0.15	0.17	66	0.04	0.02	0.04	0.00	**0.74**	0.00	**0.17**
wiki	0.07	0.89	0.96	75	**0.08**	0.04	0.02	0.02	0.01	**0.32**	**0.52**
web	NA	92.27	92.27	3354	0.04	0.04	0.02	0.01	0.00	0.02	**0.87**

Table 3. Left: Percentages of classified clicks into the different intents with the rule-based "Rule", the machine learning "SVM" methods, and combined together "Both". **Right:** Percentage of correctly classified/ mis-classified intents. **N** is number of training examples (values above 0.05 in bold).

3.1 Rule-based classifier

Our aim here is to classify as accurately as possible some of the log entries. This was important as these classifications were to be used as inputs to calculate the features for the machine learning approach. To build a high precision classifier, we used the most reliable source of evidence, namely the query terms. For example, if the user has explicitly typed *photo* in the query, we deduced that the user is looking for image, i.e. has an image intent. Here *photo* is referred to as an "intent-specific" term. We use the following intent-specific terms to classify the clicks into domains/genres:

Image: *image, images, photo, photos, picture, pictures.*
Video: *video, videos, movie, movies.*
News: *news.*
Map: *map, maps.*
Blog: *blog, blogs.*
Wikipedia: *wiki, wikipedia.*

Although not exhaustive, these intent-specific terms are a good approximation of how users would search, in terms of the queries they would submit to a search engine, for documents of particular domains or genres. With this approach, 268,491 log entries (i.e. clicked URLs), that is 4% of the total log entries, were identified to be of one of the six intents.

3.2 Machine learning classifier

To identify additional non-web intents, we make use of machine learning techniques. This requires (1) a manually labelled set of log entries; (2) designing features correlated with the possible intents that can be computed for every log entry; (3) training a classifier to predict the intent given (1) and (2). Finally, we predict the intent of the non-manually classified log entries using the classifier given their features.

For (1), we randomly sampled 3800 log entries and manually classified them into one of the following six *category* intents – image, video, blog, map, wikipedia, and news. A click was classified as having a web intent (our unclassified category) when it could not be classified into any of the above six categories. A web intent corresponds mostly to the typical web search result, and we expect it to be the predominant intent. We used the query terms, the URL and the document title for this purpose. The outcomes of the manual classification is shown in Table 3, the **N** column. As expected, web corresponds to the predominant query intent.

Step (2) aims at defining a set of features associated with each log entry, where each feature should be a good predictor for identifying a subset of intents (ideally one). In our case, the set of features were based on the language models computed from the dataset (classified intents) obtained through the rule-based approach. We build a language model for each source of evidence and each category. We also build a background language model for the source of evidence itself, which estimates the probability of a term to appear in the given source regardless of the intent. We chose to model separately the three sources of evidence since they are of very different nature.

Our hypothesis is that each category uses a vocabulary often associated with explicit intent-specific terms. For instance, in a query, if "Aniston" is often associated with "photo", the term "Aniston" will be associated to a large number of log entries classified as an image intent by the rule-based classifier. As a result, the query language model for the intent "image" will give a higher probability to "Aniston" than the background language model, and thus comparing both probabilities gives the classifier an indication of how likely a term (or a set of terms) is generated by an intent-specific language model rather than by the background one. We estimated the parameters of each language model[2], one for each source of evidence $\mathbf{s}$ and intent $\mathbf{i}$ (21 in total, i.e. one for each of the 3 sources and for each of the 6 intents plus the background model). We estimated the probability that a term t occurs using the standard maximum likelihood estimate, and smoothed it using the background language model for a given source of evidence.

$$P(t/\mathbf{i},\mathbf{s}) = \lambda P_{ML}(t/\mathbf{i},\mathbf{s}) + (1-\lambda)P_{ML}(t/\mathbf{s}) \quad (1)$$

$$= \lambda \frac{c_{\mathbf{i},\mathbf{s}}(t)}{\sum_{t'} c_{\mathbf{i},\mathbf{s}}(t')} + (1-\lambda)\frac{c_{\mathbf{s}}(t)}{\sum_{t'} c_{\mathbf{s}}(t')} \quad (2)$$

[2] For the URLs, we considered that terms were any maximal sequence of alphanumeric characters. For example, `http://www.abc.com/video` has four terms, www, abc, com and video.

The probability P_{ML} is the maximum likelihood estimate of the probability of a term occurring in a given source of evidence $\mathbf{s}$, and if given, for the intent $\mathbf{i}$ (otherwise, it is the background language model). Here, $c_{\mathbf{i},\mathbf{s}}(t)$ denotes the number of times the term t appeared for source $\mathbf{s}$ with the intent $\mathbf{i}$, and $c_{\mathbf{s}}(t)$ is the number of times term t appeared for source $\mathbf{s}$. These were computed from the set of automatically classified clicks using the rule-based classifier. The smoothing parameter λ was heuristically set to 0.95, to emphasize the importance of the intent. We then compute the probability that a sequence of terms T is generated by any of the language models by

$$P(T/\mathbf{s},\mathbf{i}) = \prod_{t \in T} P(t/\mathbf{s},\mathbf{i}) \tag{3}$$

We then use the logarithmic ratio of the probability (for a given source) of observing T given the intent to the probability of observing T:

$$R_{\mathbf{i},\mathbf{s}}(T) = \log \frac{P(T/\mathbf{s},\mathbf{i})}{P(T/\mathbf{s})} \tag{4}$$

whose value is above 0 if it is more likely that the text was generated given the intent than in general, and below 0 in the opposite case. This gives rise to a set of 18 features (one for each of the 6 categories and 3 sources) that are used as an input to build a multi-class classifier.

We use an SVM classifier[3] [1] because it performed the best when evaluated with a 10-fold cross-validation, described further below (using nine tenth of the manually classified data to learn, and one tenth to compute the performance, and repeating this operation 10 times). During the selection process, we preferred models that predicted either the correct or the web intents, over those with better performance that predicted an incorrect non-web intent.

The SVM classifier was then trained using a 3-fold cross validation and a Gaussian radial basis function. With respect to the manually labelled data, to give less importance to the web intent, we down-sampled the number of corresponding examples and only chose 20% of those, which gives a total of 817 manually labelled examples. In addition, we added an equal number of automatically labelled log entries (randomly sampled among the 224,241 classified log entries), using the same rules as in Section 3.1 but using the title and URL as sources of evidence. We experimentally found that using this set of data for training did improve the performance of the classifier.

Table 3 (Right) shows the confusion matrix with our final settings. The correct classification rate is low (between 0.27 and 0.87), but most of the time, when a click is misclassified its predicted intent is web. The only exceptions are for image (18% are classified as video) and blog (20% are classified as either image or video). Nonetheless, given that web is our unclassified category, the results show that we have improved recall without hurting precision too much.

Table 3 (Left) shows the statistics about the intent classification (rule-based, using SVM as our machine learning approach, and merged together) of all the

[3] We used the implementation of [2].

log entries. It is this *merged* labelled log that was used in the remaining analysis of this paper. We see that approximately 8% of the total log entries were identified to have a domain or genre intent other than web. This is not negligible considering the large size of the log data analyzed, and is compatible (although not directly comparable) to the results reported in [5].

4 Research questions and methodology

To study the domain and genre intents, and their dynamics, we posit the following three research questions, each investigated in separate sections next: (**R1**) What are the frequent combinations of domain and genre intents within a search session? (**R2**) Do domain and genre intents evolve according to some patterns? and (**R3**) Is there a relation between query reformulation and a change of intent?

From Table 2, column "Original", around 96.48% of the sessions have only one underlying intent. However, most sessions are composed of one or two clicks (85%, computed from Table 1), and web is the most likely intent (92.28% in Table 3 Left). We can expect that a high percentage of sessions will be one- or two-click sessions with a web intent, and hence will be single-intent sessions. It is therefore not possible to know whether our statistics of 96.48% is due to the fact that users do not combine intents, or that the click and intent distributions are highly skewed. To overcome this, we compare statistics with those obtained from a *random log*. This random log is exactly similar to the real log, but instead of using a classifier to assign an intent to each log entry we have to select one by random, in accordance with the intent distributions presented in Table 1. The random log is a log where the intent would be independent of what the user is searching for, and of his or her search history. To compute statistics for this random log, we average over all the possible random intent assignments.

We thus report statistics computed for both the real and the random logs. Going back to the example of the 96.48% of single-intent sessions in Table 2, we can see (Table 2, column "Random") that we would expect 93.14% of the sessions to be single-intent in the random log. This sets the following limit: if the real number was below this limit (even with a percentage as high as 90%), then we could say that users tend to combine more than one intent within a session; instead, we observe that our statistic is higher (96.48%) which means that sessions are indeed generally single-intent.

5 Combination of query intents

We investigate the existence of frequent combinations of query intents in search sessions (**R1**). We compute how often query intents, as classified with the approach described in Section 3, co-occur within the same search session. Table 2 shows the percentage of sessions that contain two or more different intents. We observe that there are very few sessions with more than two different query intents. This is in accordance with the study reported in [5].

Table 4. Percentage of sessions with corresponding pair of intents, where L stands for original log, and R for random log, n= News, m= Map, i= Image, v=Video, w= Wikipedia, b=Blog and W= Web. Column % 1/2 (respectively % 2/1) reports the percentage of sessions with the first intent of the pair (respectively second) that also had the second (respectively first) intent.

	%		% 2/1		% 1/2	
Comb	**L**	**R**	**L**	**R**	**L**	**R**
bm	0.00	0.02	0.5	4.6	0.1	0.5
nb	0.01	0.03	0.3	0.5	2.6	6.2
vm	0.01	0.2	0.3	4.5	0.4	4.4
bw	0.02	0.01	4	2.8	0.7	0.5
ib	0.02	0.02	0.8	0.5	4.1	4.3
im	0.02	0.19	0.9	4.5	1.1	4.1
in	0.04	0.26	1.5	6.2	1.1	4.2
nm	0.04	0.28	1	4.5	1.8	6.2
nw	0.04	0.18	1.1	2.8	1.6	6.2
vb	0.04	0.02	1.4	0.5	7.8	4.5
wm	0.04	0.13	1.5	4.5	1.9	2.8
vn	0.05	0.27	1.7	6.2	1.3	4.4
iw	0.09	0.12	3.3	2.8	3.4	4.1
iv	0.10	0.18	3.6	4.4	3.5	4.2
vw	0.06	0.12	2.1	2.8	2.2	4.4

	%		% 2/1		% 1/2	
Comb	**L**	**R**	**L**	**R**	**L**	**R**
bb	0.09	0.00	18.9	0.2	18.9	0.2
ww	0.50	0.04	18.6	1.5	18.6	1.5
mm	1.11	0.1	51.8	2.3	51.8	2.3
vv	1.17	0.1	43.1	2.2	43.1	2.2
nn	1.33	0.2	35.5	3.2	35.5	3.2
ii	1.40	0.09	52.5	2.1	52.5	2.1
bW	0.43	0.51	88.6	97.1	0.4	0.5
mW	1.35	4.42	62.7	97.3	1.4	4.4
iW	1.85	4.09	69.2	97.3	1.9	4.1
vW	1.99	4.31	73.1	97.3	2.1	4.3
wW	2.48	2.77	92.9	97.2	2.6	2.8
nW	2.87	6.1	76.5	97.4	3	6.1
WW	91.77	91.41	94.9	91.7	94.9	91.7

The ratio between the original and the random log statistics show that the fact that a session is associated with a low number of intents is not due to chance. Moreover, the ratio increases as the number of intents increase, which shows that when users have diverse intents, it is generally restricted to at most two. Therefore, we computed the percentage of sessions where at least two intents appeared among sessions with two intents or more (Table 4). For instance, the value 0.01% for "nb" means that there are very few sessions with a blog and a news intents. We also computed for each pair of intents the % that given one intent the second was observed within the same session, and vice versa (3rd and 4th group of columns). For instance, for "vw", the value 2.1% means that 2.1% of the sessions that had a video intent also had a wikipedia intent.

We can observe that most users do not mix intents. Indeed, rows "bm" to "vw" and rows "bW" to "nW" show that users are less likely to combine two different intents in the same session than what would be expected by random (around 3 times less likely in average). Looking at the rows "bb" to "ii", it is on average around ten times more likely that users repeat a click on the same intent than what would be expected by random. In sessions made of two or more clicks, when one intent is map, video, image or news, then there is above 35% of chance to observe a second click with the same intent (third and fourth group of columns). For blog and wikipedia, the probability is lower although still high (around 19%). This could be because users might consider wikipedia and blog result pages as web pages and hence do not differentiate them as belonging

to different categories. We however observe some potential exceptions for the pairs blog/wikipedia ("bw"), image/blog ("ib"), and video/blog ("vb"). These three pairs occur more often together that would be expected by random. However, the difference in percentages (≈ 0.01) are so low that this is likely due to noise from the classification.

A last observation is that when there are two intents, these are often a web intent and any non-web intent, as shown in the last series of rows in the table. This is not surprising, and means that search results should continue to contain mostly web results, and when appropriate, images, videos, blogs, etc in addition. This is nowadays the approach followed by all major search engines.

6 Patterns of query intents

We investigate how domain and genre intents evolve within search sessions (**R2**). We restrict ourselves to sessions with two query intents. We consider the five most frequent co-occurrences of two query intents (other co-occurrences were too low): image+web, wiki+web, video+web, news+web, map+web. For each such pair intent+web, we looked at all possible sequences of changes of query intents. The four most frequent ones, for all five pairs, were of the form, web $\rightarrow$ intent, intent $\rightarrow$ web, intent $\rightarrow$ web $\rightarrow$ intent, and web $\rightarrow$ intent $\rightarrow$ web. In Table 5, we report for each pair the percentage of sessions containing each of the identified four sequences. All the others come under "Other". In our calculation, we excluded sessions with less than three clicks, to avoid results biased towards the large number of two-click sessions.

First, for the wikipedia intent, users do not follow any particular pattern. Indeed, the sequences obtained are close to what would be expected by random. This confirms the findings of the previous section, where we made the hypothesis that users do not differentiate between wikipedia and web documents. Second, when the intent is news, video and map, users switch from one intent to another, but do not tend to switch back to the first intent. By random we would expect more users to move back and forth between intents. This can be seen in the difference between the random and real logs for the four sequences. Third, in the case of image intents, we observe that users are less likely to move back and forth between intents. However, different from news, video and map intents, users are more likely to begin with a web intent before looking at images, rather than start with an image intent and then switching to a web intent.

There is a common sequence in the intents for all except for wikipedia. Users have a tendency to go from one intent to the other, and then to end the session, rather than switching several times between intents within the same session.

7 Query intents and query re-formulation

We study both quantitatively (how many) and qualitatively (how) the effect of a change of intent on a user query (**R3**). We thus compare pairs of consecutive queries with two different intents, within the same session. For each such pair,

Sequence	Original	Random
wiki → web	25	26
web → wiki	30	26
web → wiki → web	**41**	**46**
wiki → web → wiki	1	0
Other	3	1
image → web	27	26
web → image	**37**	26
web → image → web	31	**46**
image → web → image	1	0
Other	4	1

Sequence	Original	Random
video → web	32	26
web → video	**34**	26
web → video → web	31	**46**
video → web → video	1	0
Other	2	1
map → web	**36**	26
web → map	34	26
web → map → web	28	**46**
map → web → map	1	0
Other	1	1
news → web	**34**	26
web → news	31	26
web → news → web	32	**45**
news → web → news	1	1
Other	2	1

Table 5. Sequence of intents in search sessions, for each pair web+non-web. Percentage numbers for the most frequent sequences are in bold.

we computed the numbers of queries that were exactly the same, modified or completely different. Results are shown in Table 6.

For blog and wikipedia, over 50% of the users did not usually change their query (going from a blog or wikipedia intent to web intent or vice versa). It is likely that this happens because both types of results are present in the top ranked documents for the same query. Users do not have to change their queries to obtain results from blog/wikipedia and then web sources (and vice versa). The situation is reversed for news, image, video, and map. Most of the time, users did change their query (over 65%). We also observed that there was a slight difference between news/map/video and image intents. For the former, users issued different queries, whereas for the latter, in half of the cases, users modified their queries by adding or removing terms. We found that, for news/map/video, often users often changed their search topic (e.g. "sovereign bank center trenton" to "art of 1769") and hence modified the query completely, whereas for image users seem to have looked at the results and then added intent-specific terms (e.g., "photo").

When a query was modified, we also looked at which terms were added or removed. We easily identified terms linked with an intent, i.e. intent-specific terms; e.g. "wikipedia", "what", "how" for wikipedia; "blog", "how", for blogs;' 'news", "newspaper", and "press" for a news intent; etc. Some of these terms were present in the rule-based classifier described in Section 3.1.

8 Conclusion

We analysed a click dataset to obtain an understanding of domain (image, video, and map) and genre (news, blog and wikipedia) intents, and their dynamics.

Sequence	Exact	Modified	Different	Sequence	Exact	Modified	Different
web → blog	**56**	24	19	web → video	35	25	**40**
blog → web	**52**	23	25	video → web	33	24	**43**
web → wiki	**59**	18	23	web → image	30	**40**	29
wiki → web	**54**	20	26	image → web	30	34	**37**
web → news	21	21	**58**	web → map	4	24	**73**
news → web	18	19	**63**	map → web	3	21	**76**

Table 6. Percentage of sessions where a query was not modified, was modified (i.e. by adding or removing terms), and was different (no terms in common).

The first step was to identify the domain and genre intents behind a query. Using a rule-based and an SVM classifier, we classified approximately 8% of the total click dataset to have one of the six domain or genre intents. We looked at how intents co-occur within a session. We observed that users do not often mix intents, and if they do, they mostly use two intents. Furthermore, these were often a web intent and any of the non-web one. Second, we investigated if these intent combinations evolve according to some patterns. Our results show that, except with wikipedia, users in general tend to follow the same intent for a while and then switch to another intent. In other words, users do not switch back and forth between intents. Third, we were interested to see if there were relations between query re-formulation and change of intent. We observed that for video, news and map intents, often completely different queries were submitted, whereas, for blog and wikipedia intents, the same query was used. Further, intent-specific terms were often used when the query was modified.

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

References

1. K. Crammerand Y. Singer. On the Learnability and Design of Output Codes for Multiclass Problems. *Computational Learning Theory*, pages 35-46, 2000.
2. A. Karatzoglou, D. Meyer and K. Hornik Support Vector Machines in R. *Journal of Statistical Software*, Volume 15, Issue 9, 2006.
3. Kc.Y. He, Y.S. Chang and W.H Lu. Improving Identification of Latent User Goals through Search-Result Snippet Classification. *WI*, pages 683–686, 2007.
4. http://www.iprospect.com/about/researchstudy_2008_blendedsearchresults.htm
5. J. Arguello, F. Diaz, J. Callan and J. Crespo. Sources of evidence for vertical selection. *ACM SIGIR*, pages 315–322, 2009.
6. N. Craswell, R. Jones, G. Dupret and E. Viegas (eds). *Proceedings of the 2009 workshop on web Search Click Data*, 2009.