Towards a Deeper Understanding of the User’s Query Intent

Liliana Calderón-Benavides  
University Pompeu Fabra  
Barcelona, Spain  
liliana.calderon@upf.edu

Cristina González-Caro  
University Pompeu Fabra  
Barcelona, Spain  
cristina.gonzalez@upf.edu

Ricardo Baeza-Yates  
Yahoo! Research  
Barcelona, Spain  
rbaeza@acm.org

ABSTRACT

The identification of the user’s intent behind a web query is considered to be one of the most important challenges for the modern information retrieval systems. Although there have been studies that characterize some possible dimensions for the user intents, most of these studies just place each query into one or two dimensions. In order to approach the real world, search engines need to have a better understanding of the user’s intent and its multidimensional nature. In this paper we study, analyze and characterize a wide range of facets or dimensions that may be useful for user’s intents identification when searching for information on the Web. We outline the relationships and dependencies that exist between these dimensions, and identify the contributions of each in the task of accurately recognizing user’s intentions in Web search.

Categories and Subject Descriptors

H.3.3 [Information Storage and Retrieval]: Information Search and Retrieval—Query processing, user preferences and personalization

Keywords

Web Usage Mining, Query Analysis, User’s Intent

1. INTRODUCTION

Searching the Web for useful resources is not an easy task. Most of the users use short queries to describe their information needs. Starting from the query words, a search engine should be able to accurately infer the user’s intents and to guide the user to obtain the actual desired information.

Search Engines require the use of different strategies to meet the user’s intents. It is necessary to combine several factors in order to gain insight into the real goal of users’ search. For example, is the user searching for a document to read it?, or does he/she want to perform a transaction such as buying a product?, is it important that the retrieved information comes from trusted Web sites? –or the source of the information is not a relevant feature, as long as the content is good-. how important is the information recency?, what kind of resource is the user searching for?. what the user wants to do with such resource? –download it, read it online, or just find a reference. These questions help us to build a picture of the multi-dimensional nature of the user’s intent, and show why is it necessary to place the user’s requirements in a wider context.

1.1 Related studies

Several authors have studied Web query logs to explore the issue of defining and discovering user intent in information search on the Web. Broder [3] defined a classification of web queries into three categories: informational, transactional, and navigational. Rose and Levinson [9] introduced goal hierarchy and presented manual classification results on 1500 queries from Altavista query logs. Following Broder’s taxonomy, several authors have focused their work on the automatic classification and characterization of the user intents [5, 6, 7]. Other researchers have worked on large manually annotated data sets, such as Baeza-Yates et al. [2] where a data set of around 6,000 popular queries were classified into two aspects: intention (Informational, Not Informational and Ambiguous) and topic (ODP1 topics). Nguyen et al. [8] proposed a facet classification to further help in possible actions that may be taken by a search engine to aid the search process. The facets defined by them were: Ambiguity, Authority Sensitivity, Temporal Sensitivity and Spatial Sensitivity. Recently the interest in determining user intentions has spread to commercial [1] and geographical [4] applications.

Although previous studies—from different perspectives—have addressed, the issue of determining user intents, there is still not an approach that integrates such perspectives. Each study analyzes its own proposal, and even though the conclusions are useful and the relevance of the results is shown, a comprehensive study that explore the integration and the relation among them is needed. Hence, with the aim to provide a more comprehensive understanding of the user and his/her intents, in this paper we study and evaluate a set of facets or dimensions2 that we expect will give a more accurate representation of users needs and intents.

1Open Directory Project www.dmoz.org/
2Throughout this paper, the terms facet and dimension will be used to represent the same concept.
Table 2: Inter-annotation Agreement and Kappa Coefficient for the facets

<table>
<thead>
<tr>
<th>Facet</th>
<th>Overall Agreement</th>
<th>Kappa Coefficient</th>
</tr>
</thead>
<tbody>
<tr>
<td>Time S.</td>
<td>99.23</td>
<td>0.98</td>
</tr>
<tr>
<td>Scope</td>
<td>96.74</td>
<td>0.93</td>
</tr>
<tr>
<td>Objective</td>
<td>92.54</td>
<td>0.85</td>
</tr>
<tr>
<td>Authority S.</td>
<td>84.32</td>
<td>0.69</td>
</tr>
<tr>
<td>Topic</td>
<td>68.26</td>
<td>0.66</td>
</tr>
<tr>
<td>Task</td>
<td>75.71</td>
<td>0.63</td>
</tr>
<tr>
<td>Spatial S.</td>
<td>81.07</td>
<td>0.62</td>
</tr>
<tr>
<td>Genre</td>
<td>65.00</td>
<td>0.53</td>
</tr>
<tr>
<td>Specificity</td>
<td>55.44</td>
<td>0.33</td>
</tr>
</tbody>
</table>

Figure 1: Objective of the query with respect to genre

3. QUERY CLASSIFICATION

We used a sample of 5,249 queries from the TodoCL search engine query log. The complete set was manually classified by a group of judges. In order to estimate the reliability of the set of assessments, 10% of the queries were classified by two judges. Table 2 contains the overall agreement of the judges as well as the Kappa coefficient values.

Results from the overall agreement indicate that the consistency of the manual classification is highly satisfactory. In average, the overall agreement is 80%. Eight out to the nine facets have reached an overall agreement higher than 65%, which is quite high if we consider the number of dimensions that were assessed, the number of possible values that each dimension can take, as well as the different criteria and the subjectivity of the judges.

In order to determine the extent to which the observed agreement exceeds what is expected to obtain by chance, we used Kappa coefficient. The analysis of the values obtained from Kappa was done following the interpretation proposed by Landis and Koch (cited in [10]). In light of this interpretation, the agreement of the judges for the facets is: time sensitivity, scope and objective are almost perfect; the values obtained for authority sensitivity, topic, task and spatial sensitivity are substantial; the agreement for the facet genre was moderate; and the consensus for specificity is fair.

The values obtained from Kappa coefficient reflect a consistency and reliability of the manual classification. The assessment from the judges for all the facets is beyond the chance, and even for three facets, the agreement was almost perfect.

3.1 Correlation Among Dimensions

Aiming to establish the relations among the dimensions, we calculated the correlation factor among all of them. From the results we can outline which features are better correlated. In general, genre and objective are the dimensions that are more highly correlated. Figure 1 shows the distribution of queries between these two dimensions, revealing specific relations among its values. The strongest association is the one formed by genre=Community and objective=Resource; the genre=Reference and genre=News also show a clear association with the Resource objective. The trend of the genre associated with the objective=Resource is more oriented towards obtaining information, than to performing a specific action. On the other hand, the genre=Business exhibits a different behaviour compared to the others. genre=Business is the genre with highest number of queries with objective=Action, this genre concentrates the highest percentage of transactions –at commercial level–, this fact is reflected by the distribution of the queries.

One of the most meaningful dimensions is topic. By contrasting this dimension with the others, we observe the trends of the users’ intents. For example, when we associate topic with genre we find that some topics clearly belong to a particular genre, while others maintain a balanced distribution among the different genres.

The most important topics for each genre are: genre=Business {Finance, Cars & Transportation, Computers & Internet, Home & Garden}, genre=Community {Politics & Government, Health, Entertainment, Music & Games, Arts & Culture}, genre=News {Entertainment, Music & Games, Politics & Government, Health, Finance}, genre=Reference {Science & Mathematics, Travel, Education}. While the number of queries grouped in the
<table>
<thead>
<tr>
<th>Facet</th>
<th>Values</th>
<th>Description</th>
<th>Use</th>
</tr>
</thead>
<tbody>
<tr>
<td>Genre</td>
<td>News, Business, Reference, Community</td>
<td>It is consider a meta-category that provides a generic context to the user query intent.</td>
<td>This facet clearly delimits the search engines area of search, and is expected to make the search faster and more accurate. For example if the genre=Community, the pages that have to be looked at are blogs, forums, chats, and communities, such as Facebook.</td>
</tr>
<tr>
<td>Topic</td>
<td>Adult &amp; Sex, Arts &amp; Culture, Beauty &amp; Style, Cars &amp; Transportation, Computers &amp; Internet, Education, Entertainment, Music &amp; Games, Finance, Food &amp; Drink, Health, Home &amp; Garden, Industrial Goods &amp; Services, Politics &amp; Government, Religion &amp; belief systems, Science &amp; Mathematics, Social Science, Sports, Technology &amp; Electronic, Travel, Undefined, Work</td>
<td>This list of topics was created from the first level of categories offered by ODP, Yahoo!, and Wikipedia, but leaving out categories that are actually a genre such a News and Reference.</td>
<td>Although this is one of the first features that are considered by the search engines, they should ideally present the answers to the user organized by topic</td>
</tr>
<tr>
<td>Task</td>
<td>Informational, Not Informational, Both</td>
<td>Considered as the primary need reflected by a query, could be also the intent itself [2].</td>
<td>If the query is Informational, then Web resources to be recommended should avoid those with transactional content, or the ones where the user has to assume an active role of interaction, such as fulfilling a form (i.e., Not-Informational resources).</td>
</tr>
<tr>
<td>Objective</td>
<td>Resource, Action</td>
<td>Represents the aim of a query, without considering the format of the information to retrieve.</td>
<td>If the user wants to perform an Action, then he/she may also be interested in commercial sites that offer a product or service, hence appropriate ads may be presented to the user.</td>
</tr>
<tr>
<td>Specificity</td>
<td>Specific, Medium, Broad</td>
<td>This facet describes how specialized is a query.</td>
<td>The search engines can help the user by presenting many similar pages, for example on the same topic, or a broad variety of pages.</td>
</tr>
<tr>
<td>Scope</td>
<td>Yes, No</td>
<td>Shows whether the query contains polysemic words or not.</td>
<td>The answers presented to the user by the search engine should ideally be presented by topic, and trying to cover most of the topics related to the polysemic words, allowing the user to select the right one.</td>
</tr>
<tr>
<td>Authority Sensitivity</td>
<td>Yes, No</td>
<td>Is the query designed to retrieve authoritative and trusted answers? [8]</td>
<td>Through this dimension, search engines have the opportunity to be very selective regarding the authority of the web pages presented to the user. For example, it is critical to trust the pages where an antivirus software would be downloaded from. Moreover, it is essential to reach the right page for all the navigational queries.</td>
</tr>
<tr>
<td>Spatial Sensitivity</td>
<td>Yes, No</td>
<td>Reflect the interest of the user to obtain a resource that is related to a particular spatial location (explicit or not).</td>
<td>Although a query does not mention a place, search engines should be able to identify spatial sensitivity to offer information related to the local area from which the query was submitted, emphasizing in queries that reflect a commercial need. For example: “school”, “tai chi classes”, “cars selling”.</td>
</tr>
<tr>
<td>Time Sensitivity</td>
<td>Yes, No</td>
<td>Whether the information to retrieve involves a date or period of time.</td>
<td>Search engines can take advantage of the time sensitivity of queries to suggest related information that occurred in the same period as what it is being sought for. Further more, this information can be used to recommend resources following a temporal order, as it is done with news.</td>
</tr>
</tbody>
</table>

Table 1: Description of the Facets, their values and usefulness
genre=News is not so high, topics with greatest concentration of queries have characteristics that are very representative of these genre. For example, the most important topics for the News genre are Entertainment, Music, & Games and Politics & Government, which are topics that are continuously changing and generating new information to a wide community.

The genre facet complements topical associations with an idea about what the user wants to do in a specific topic area. It is different to know that the query “hp printers” belongs to the topic=Technology and Electronic, than to know that the same query belongs to the topic=Technology and Electronic as well as to the genre=Business. Furthermore, following with the last example, if we find that the facet objective=Action then, there is a very high probability that the user intent is to perform a commercial transaction, in the topic=Technology and Electronic.

Dimensions such as scope and specificity are correlated too. We found that queries with specificity=Broad were assessed to have polisemic words too, i.e., scope=yes. The dimensions that represent a kind of sensitivity: author, spatial and time are related among them. There is an association between time and the other two sensitivities. Despite that, in the complete data set, the number of queries labelled with time sensitivity=yes is not very high, we found that when this dimension is present in a query, it is usually accompanied by another kind of sensitivity. Usually, when a query is specific enough, that is, it includes a concrete date or a particular time space, the query is also specific enough to include another kind of sensitivity.

4. DISCUSSION AND FUTURE WORK

Web Search Engines tend to view the query formulation and the retrieval process as a simple task-focused activity. However, the intentions behind Web queries are more complex and the search process should be more oriented to a variety of characteristics than only to a simple task. In this paper we have studied, analyzed and characterized a wide range of facets or dimensions that can be useful for user’s intent identification when searching for information on the Web. These dimensions/facets are: genre, objective, specificity, scope, topic, task, authority sensitivity, spatial sensitivity and time sensitivity.

From the analysis of the manual classification of queries we found that dimensions such as scope, topic, or objective are easier to determine than genre and task. In order to classify a query, the former dimensions require a lower level of subjectivity from the judges, in comparison to the latter ones.

We have confirmed the benefit to separate specific topics, from those that are more general purpose. With the topics that are general purpose we have created the genre dimension, which in turn is considered as a meta-dimension. The use of this meta-dimension allows a fine-grained classification of queries at the topic level, as well as a more direct relation with the objective pursued by the user. For example, regardless the topic of a query, the genre=Business, in most of the cases, is related to perform an Action, which is one of the values of the objective dimension.

This is an ongoing work, hence the results presented in this paper are neither complete, nor definitive. We will continue analysing the correlations among the dimension/facets in order to filter out the less representative ones, while keeping the most descriptives. Our aim is to provide an accurate model of the real world of needs and intents of users. This work is the basis to determine and classify the user’s intents in an automatic way.

5. REFERENCES