A multi-dimensional model for search intent

Max Hinne
Institute for Computing and Information Sciences (iCIS)
Radboud University Nijmegen
mhinne@sci.ru.nl

Maarten van der Heijden
Institute for Computing and Information Sciences (iCIS)
Radboud University Nijmegen
m.vanderheijden@cs.ru.nl

Suzan Verberne
Centre for Language and Speech Technology
Radboud University Nijmegen
s.verberne@cs.ru.nl

Wessel Kraaij
Institute for Computing and Information Sciences (iCIS)
Radboud University Nijmegen
TNO, Delft
kraaijw@acm.org

ABSTRACT

The interaction of users with search engines is part of goal driven behaviour involving an underlying information need. Information needs range from simple lookups to complex long-term desk studies. This paper proposes a new multi-dimensional model for search intent, which can be used for the description of search sessions. Using examples from a search engine log we show that our model allows a more comprehensive description of information need than existing categorizations.

1. INTRODUCTION AND BACKGROUND

User interaction with search engines is an object of study in different domains of science. This may be the reason that key concepts such as *intent, information need* and *query session* lack a consistent definition in the literature. Many definitions of query sessions have been suggested and explored in the literature [4], It seems well accepted that sessions can consist of multiple queries that are often topically related. Gayo-Avello [2] introduce the term *searching episode* for all queries by a user during a single day, consisting of one or more *search sessions* where the "successive queries are related to a single information need or goal". Session boundaries are usually determined by looking at lexical or temporal cues or a combination of these cues.

Classifications of search patterns that can help to determine session boundaries have been presented in e.g. Lau and Horvitz [5] and He et al. [3]. A key element of the search patterns within a search session is that there is some form of lexical overlap. Queries can be refined by specialization, generalization or reformulation. These refinement classes are examples of what Lau and Horvitz call *user's intents relative to his prior query*. Thus in an IR context, intents could be defined as intermediate goals that are the result of a certain knowledge state, which is the result of the interaction with the search engine so far. Intents represent the (sub)goals motivating the user's search behaviour.

We introduce a multi-dimensional notion of intent, with information need as the driving force behind search behaviour, and search intent as specializations of that force.

Since information need is an abstract concept, it is not necessarily restricted to a specific search session. This aspect is important, since the overall information need is a core part of the context that can help to define the relevance of search results. If a search engine can detect that e.g. a request for booking skiing lessons is connected to a previous search session concerning renting an apartment in a specific ski resort, it would be helpful to rank the pages about ski-schools in the vicinity of this ski-resort higher than pages about other ski-schools.

In this paper, we will show that such a multi-dimensional view on intent can be supported by click data. We propose three facets of search intent, explained in Section 2. We claim that these facets can help to create a more fine grained taxonomy to discuss and analyze search intent. We are also able to relate several existing intent classification schemas (e.g. Broder [1], Lau and Horvitz [5]) to our model (Section 3 and Section 4). Section 5 provides some examples from data followed by some concluding remarks and future work in Section 6.

2. OUR MODEL FOR INFORMATION NEED AND SEARCH INTENT

Following survey studies such as [2] and [7] , we conclude that the concepts *information need* and *search intent* (or *query intent*) are widely used in the literature about user interactions with search engines, but lack a uniform interpretation. Before we discuss extensions to existing classification schemes for search intent, we present our view on the relation between information need and search intent:

At the basis of a user interaction with a search engine lies the *information need*. This can be anything from an abstract, unexpressed need to a clearly formulated request. A complex information need generates one or more *search intents*. A search intent is a clear-cut element of the information need that the user hopes to solve with a well-formulated query. In practice, a search intent leads to the realization of one or more queries; it is possible that a user needs to formulate multiple queries until the local search intent is satisfied. In that case multiple queries are related, motivated by the user's desire to refine a query.

This hierarchical process, starting at an information need

DIR'2011, February 4, 2011, Amsterdam.
.

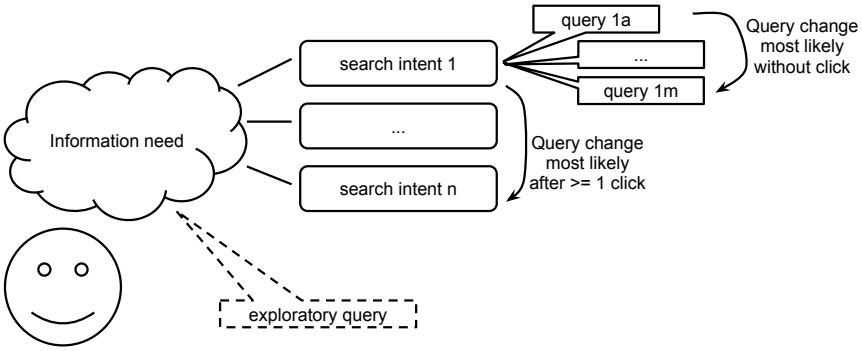

Figure 1: Our model for information need and search intent

and ending with a series of queries, is visualized in Figure 1.[1] The process can be exemplified by the following case: Consider the complex information need "Collecting information about the Dutch prime minister for an essay". It is composed of several search intents: finding out who the Dutch prime minister is, collecting biographical facts about Mark Rutte, finding a good picture, and foraging opinions and media performances related to him. These search intents may require multiple queries to be satisfied, and perhaps the user has to reformulate his queries multiple times in order to obtain a useful result.

3. CLASSIFICATIONS OF SEARCH INTENT

The search intents generated by an information need are traditionally classified according to the *actions* the user wants to execute with the results. These can be *informational, transactional* or *navigational* [1]. We argue that although this classification is sound, it is not complete.[2] It forms one dimension of the three-dimensional classification of search intent that we propose in this section.

Sushmita et al. [8] propose that search intents should be classified by the requested *form* of the results. For example, a search may be aimed at retrieving pictures, maps, videos or Wikipedia entries. They refer to this aspect of the search intent as a combination of *query domain* and *query genre*. We will instead use the term *mode* to refer to this second dimension of the search intent. The user's choice along this dimension is sometimes made explicit in the query, by adding terms such as "pictures" or "movies".

The third dimension that characterizes the search intent is its *topic*. This is most strongly connected to the textual realization of the query: the query "Mark Rutte" is a request for items 'about' Mark Rutte. Within one session of interaction with a search engine, the user may consider multiple topics, that each relate to a series of queries.

In most papers addressing information need, queries are classified according to the search intent that generated them, using the navigational-transactional-informational scheme. We propose to extend this scheme to a three-dimensional classification, of which the axes are *action*, *mode* and *topic*. In the remainder of this paper, we investigate the relevance and applicability of these dimensions by considering series of queries in search engine interactions. We use search engine log data for this purpose, the Microsoft "Accelerating Search in Academic Research Spring 2006 Data Asset", which contains one month of MS search queries from the spring of 2006 together with the URLs clicked, a timestamp and a session identifier. Because of privacy concerns, session lengths have been cut-off at 30 minutes.

4. CLASSIFICATION OF QUERY TRANSITIONS

The usefulness of the additional dimensions for query classification become apparent when we consider the transitions between different queries. In the three-dimensional model, we expect a new query within a user session to change on one or more axes of the model. Therefore, in this section, we study the *transitions* of one query to another within one session and try to classify these transitions according to the multi-dimensional model of search intent proposed above. From our hierarchical model of information need, it follows that there are three levels on which a query transition can take place:

1. Starting to work on a new information need.
2. Introducing a new search intent within the same information need.
3. A query reformulation (correction) for the same search intent.

When a user moves from one search intent to another, then he will reformulate the query along one of the three axes of search intent *action*, *mode* or *topic*. In other words, the change of intent is realized as a query transition. Here, the query transition categorization as proposed by Lau and Horvitz [5] can play a role. Lau and Horvitz classify query transitions according to the change in surface form (textual content) of the query, labelled as generalization, specialization, reformulation etc.

The change in surface form does not have a direct link to the change in search intent, but categorizing the query

[1]If the user has an information need that he is not able to directly express in the form of a clear search intent – what Taylor refers to as the *visceral* information need [9] – the user may generate an exploratory query. The results that are presented to the user help him in formulating his search intent.

[2]In addition, navigational search intent seems are more aimed at bypassing a browser's address bar than to actually find information, but that is not an issue that we address in the current paper.

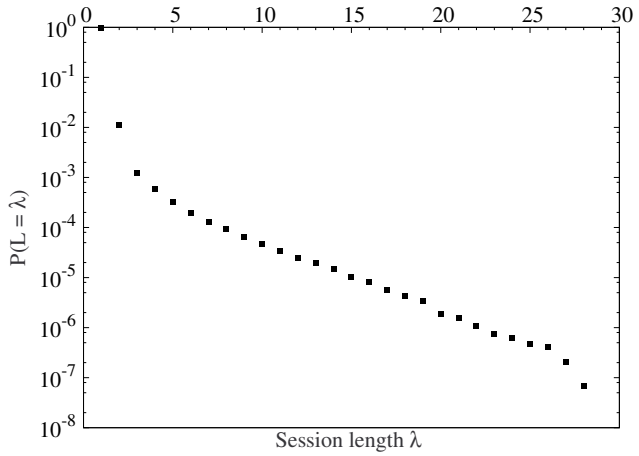

Figure 2: The distribution of session lengths in the MSN query data set as the probability for a session to contain λ clicks.

transitions may be helpful for understanding the changing intent.

We analyzed query transition behaviour with the aid of the Microsoft search log. The distribution of session length (measured in the number of clicks per session, see Figure 2) shows that 1.4% of sessions contain more than one click.[3] We implemented an automatic classification of query transitions for the MSN click data, using the following heuristics for transition types based on [5]:

- Request for additional results: query Q_{i-1} is equal to Q_i (the query is not necessarily reissued, multiple clicks for a single query show up the same way in the query log).
- Generalization: query Q_i is a substring of Q_{i-1}. E.g. "Mark Rutte prime minister", followed by "Mark Rutte".
- Specialization: query Q_{i-1} is a substring of Q_i. I.e. "Mark Rutte", followed by "Mark Rutte prime minister".
- Reformulation: query Q_i has at least one word in common with Q_{i-1} without the transition being generalization or specialization. E.g. "Mark Rutte Netherlands" followed by "Mark Rutte pictures".
- New topic: query Q_i has no words in common with Q_{i-1}.

These heuristics are oversimplified as a model for query transition because they consider queries as sequences of words that are compared literally. As a result, coincidental word overlap between queries Q_{i-1} and Q_i (such as repeating the word 'the') is categorized as a reformulation instead of a new topic. And two queries that are very similar in meaning but use different wordings (e.g. when 'pictures' is changed to 'photos') are categorized as a change to a new topic. A better implementation of the query transition categorization would be to take into account semantic relatedness between two queries. We will implement this in the near future with the use of the WordNet Relatedness tool [6].

[3]This number is quite low. It may partly be caused by the artificial cut-off of search sessions.

Table 1: 2 Million queries from the MSN click data set automatically classified into the query transition classification by [5].

Number of queries	2000000	100%
Number of sessions	1008656	
Number of follow-up queries	991344	
New topic	1339910	67.0%
New topic in same session	331254	16.6%
Request for additional results	270958	13.5%
Reformulation	247033	12.4%
Specialization	98585	4.9%
Generalization	43514	2.2%

We applied the heuristics-based classification of query transitions to the MSN click data set. In this way, all queries in a session are automatically annotated with transition information. The counts over 2 Million queries are shown in Table 1. The transition types do not explicitly inform us on the user's search intent. We argue that our suggested multi-dimensional search intent model can aid in explaining the different query transitions within a session in terms of query intent. In the next section, we use a number of examples from the click data to manually classify query transitions along the axes of our model.

5. EXAMPLES FROM CLICK DATA

We manually analyzed a number of the annotated sessions in order to gain insight in the type of transitions occurring in the data and how they relate to presumed search intents. Table 5 shows two example sessions from the click data, automatically annotated with transition information. Before analyzing this sequence of queries, we should note that since this is a retrospective analysis, the actual intents are unknown, and the analysis just shows how our model can be applied to user behaviour data.

The first three queries (0, 1, 2) in the first session seem to be informational queries about specific event locations, presumably known to the searcher (a manual check shows that query 0 leads to a restaurant chain and 1 and 2 to venues that advertise themselves as wedding reception locations). Then with query 3 the search intent seems to change, asking about wedding reception locations in a particular town in Texas, followed by a generalization in query 4. This query could be interpreted as a request for a different *mode*, i.e. a map of Seguin. Query number 5 seems to be a reformulation continuing the informational intent of query 3. Query 6, although still topically related, deals with a different facet of the information need, specifically the average cost of a wedding. After apparently finding such an estimate, the new search intent in query 7 includes the modifier 'cheap'. The last query is a specialization to the specific location 'Austin'.

Thus, the overall information need of this session seems to be about planning a wedding reception, with search intents changing to reflect different topical aspects (location and cost); different modes (information and maps); and possibly once a satisfactory location is found, the action intent might change from informational to transactional. This example thus shows that our model at least allows for a more fine grained analysis of search intents: subsequent queries can belong to different search intents while having the same underlying information need.

Table 2: Example sessions from click data, automatically annotated with transition information according to the model by Lau and Horvitz [5].

0:The Salt Lick	New topic
1:Texas Old Town Kyle , TX	New topic in same session
2:Old Glory Ranch	Reformulation of query 1 (words overlapping: Old)
3:Seguin wedding receptions	New topic in same session
4:Seguin, TX	Generalization of query 3 (words overlapping: Seguin)
5:Reception Site in Seguin, Texas	Specialization of query 4 (words overlapping: Seguin)
6:Average Cost of a wedding with 150 guests	Specialization of query 3 (words overlapping: wedding)
7:Cheap Texas Weddings	Specialization of query 5 (words overlapping: Texas)
8:Austin, Texas Wedding sites	Specialization of query 7 (words overlapping: Texas Wedding)

0:ceramic paint	New topic
1:color chart	New topic in same session
2:paint color chart	Specialization of query 1
3:paint color chart	Request for additional results (same as query 2)
4:ceiling paint that will not allow water spots	Reformulation of query 3 (words overlapping: paint)
5:ceiling problems	Reformulation of query 4 (words overlapping: ceiling)
6:water repellant ceiling	Reformulation of query 5 (words overlapping: ceiling)
7:no water stane ceiling	Reformulation of query 6 (words overlapping: water ceiling)
8:no water stain ceiling	Reformulation of query 7 (words overlapping: no water ceiling)

Let us consider an additional example, shown in the bottom half of Table 5, to gain some feeling for the classifications that our model allows. Query 0 introduces a topic as start of the session, with a query that appears *informational* and given the usual mode of internet search, *textual*. Then, with query 1 a transition is made not only in the *topic* dimension (from 'paint' to 'color') but also in the *mode* dimension, as a chart is requested. Queries 2 and 3 combine the first topics. A slight topic shift is introduced in query 4, which gives a specialization of what the paint is needed for, followed by a number of reformulations that appear to be aimed at satisfying the same search intent on water stains (one of which is just correcting a spelling error).

Again we see that a session of queries has a single information need, that is, finding information about paint that can cover water stains. Although all queries can be called informational, the topics do change from looking for ceramic paint, to colours and to paint specifically well suited to cover water stains. Queries 1–3 also clearly request a mode of information that is different from text, which we would be unable to express in Broder's classification of intents.

6. CONCLUSION AND FUTURE WORK

In this paper we proposed a multi-dimensional model for search intent. It combines three classification schemes that form its axes, viz.: the *topic* of the query; the *action* that the search results should aid in and the *mode* in which the search results are expected. A change in search intent leads to a change in query text. As a result, the changes in query texts can provide information on how the search intent of the user changed. We automatically annotated 2 Million queries from an MSN click data set with query transition classifications. We performed a manual analysis on examples of annotated sessions, showing how our model can be used to describe user search behaviour.

The added complexity of the model makes it better suited to model real data. On the down side, however, the complexity of the model makes validation more difficult. It is hard to recover what a user's search intent was, based on nothing more than the click data. As a consequence there is currently no hard validation that our model indeed captures the necessary aspects of information need and search intent. However we do believe that our more fine grained approach is valuable in understanding user queries.

In future research we will (1) make the query transition classification more informative by taking into account the semantic relatedness between subsequent queries; (2) investigate human agreement on the classification of query transitions into search intent (human agreement is a good proxy for the complexity of the problem for automatic analysis); (3) conduct a user study in which we will ask search engine user to categorize their queries in retrospect. We expect that this will provide insights in the structure search sessions and the several types of query reformulations in relation to the underlying intents.

7. REFERENCES

[1] A. Broder. A taxonomy of web search. *SIGIR Forum*, 36(2):3–10, 2002.

[2] D. Gayo-Avello. A survey on session detection methods in query logs and a proposal for future evaluation. *Information Sciences*, 179:1822–1843, 2009.

[3] D. He, A. Göker, and D. J. Harper. Combining evidence for automatic web session identification. *Information Processing and Management*, 38:727–742, 2002.

[4] B. Jansen, A. Spink, C. Blakely, and S. Koshman. Defining a session on Web search engines. *Journal of the American Society for Information Science and Technology*, 58(6):862–871, 2007.

[5] T. Lau and E. Horvitz. Patterns of search: analyzing and modeling Web query refinement. In *UM '99: Proceedings of the seventh international conference on User modeling*, pages 119–128, Secaucus, NJ, USA, 1999. Springer-Verlag New York, Inc.

[6] T. Pedersen, S. Patwardhan, and J. Michelizzi. WordNet::Similarity — Measuring the Relatedness of Concepts. In *Proceedings of the National Conference on Artificial Intelligence*, pages 1024–1025. AAAI Press, 2004.

[7] F. Silvestri. Mining query logs: Turning search usage data into knowledge. *Foundations and Trends in Information Retrieval*, 4(1-2):1–174, 2010.

[8] S. Sushmita, B. Piwowarski, and M. Lalmas. Dynamics of genre and domain intents. In *Proceedings of The Sixth Asia Information Retrieval Society Conference*, 2010.

[9] R. S. Taylor. The process of asking questions. *Amer. Doc.*, 13(4):391–396, 1962.