Web Information Retrieval for Complex Not-Informational Intents

An Introduction to User Intent Analysis

Debora Donato

Yahoo! Labs, Sunnyvale, CA USA,
debora@yahoo-inc.com

Abstract. The World Wide Web has been showing an incredible capacity of renewing itself not only by adapting in order to fulfill mutable users' needs but also fomenting new types of exigencies and requirements. For such a reason classical web information retrieval models developed around the concept of query seem not anymore adequate to satisfy complex and transactional needs for which the final goal is to accomplish a task rather than to find information. Transactional need satisfaction is not reached by showing the user with a list of documents but by reducing the total time from the moment the user issues the query to the moment the transaction is accomplished. Better support for complex queries can be obtained by a careful User Intent Analysis. In the rest of this paper, we present the reader with some of the most promising lines of research that are currently try to investigate intents and goals by focusing on all the activity related to intent satisfaction rather than on a single query.

Keywords: Web Information Retrieval, Search Engines, Users' intents

1 Motivations

Since their first appearance, web search engines have developed their retrieving models considering as input a single query and as output a list of pointers to documents relevant for the query. Such retrieval paradigm was motivated by the informational nature of the original Web, defined as the collection of the hyperlinked documents accessible through Internet.

In the last decade, we have observed a progressive shift of a number of human activities from the real life to the online world. Web sites are virtual places where people socialize, chat, play and perform a wide range of activities like bank transactions, shopping, event/travel booking and even voting. Even if informational queries are still the most frequent, transactional intents have more and more often motivated the queries issued to search engines.

As the results of the process described above, users intents have become more complex with the side effect that a single query is hardly able to capture and express all the possible facets of user needs. As an example let's consider the sequence of actions performed by a user who wants to buy an apartment. Since the

complexity of the task, the user is likely to submit a set of semantically related queries over a long temporal window; she will click on a high number of links in the results set of each submitted query with the aim of comparing different offers, browsing photos, searching for public services (like bus and schools) in the neighborhood of each of the apartments that have captured her interest. In this scenario a single query can not express all the different but related aspects behind the intent of "buying an apartment".

In this scenario, it is becoming urgent to study all the activities related to user satisfaction in order to modeling user behavior and understanding which "patterns" are more likely to lead users to success.

2 User Intent definition

User intents modeling has been a topic of interest for the last few years, but to the best of our knowledge, there is no work that tries to formalize the definition of *intent*. Most previous work [3], [12], [6], [7], [9], [10] presents automatic methods to classify query intents as informational, navigational, or transactional. According to the taxonomy introduced by Broder [3] a query is considered *i) informational* if the need behind the query is to find the document(s) that contains the desired information; *ii) navigational* if the intent is to find a particular web site; *iii) transactional* if the intent is to perform some Web-mediated activity.

As a matter of fact, partitioning the set of all queries in these three broad categories does not offer any deep insight that can be leveraged in order to better support users in their search activity.

From a qualitative point of view an intent $\mathcal{I}$ is comprised by:

- the object(s) $\mathcal{O}$ of intent;
- the verb $\mathcal{V}$, i.e. the action that the user want to perform on the object;
- a set of parameters $\mathcal{P}$ or inputs for the action.

As an example let consider the transactional query `ticket from Rome to Milan`. In this case $\mathcal{O}$ = `ticket`, $\mathcal{V}$ = `booking/buying/purchasing`, $\mathcal{P}$ = `Rome, Milano`.

As for the query we just consider, the action is often implicit. A particular case is offered by informational queries where the implicit action is always `find/read`. Intents like purchasing an house result in a set of related queries and hence in a set of objects and verbs.

Search engine users submit queries to address information needs. The expression *physical session* is used to address all the activity of a user interacting with a search engine within an inactivity interval (often set to 30 minutes). Within a single physical sessions users perform many tasks. A task or information need results then in subsequence of queries, called *logical sessions*.

Jones and Klinkner [8] break tasks into two groups: (1) *goals*, which consist of atomic information needs, and (2) *missions*, which consist of one or more goals. A typical mission is the activity needed for planning a trip, where single goals are "booking flight tickets", "booking hotels", "compiling a list of points of interest". In [8], the authors introduce a method of automatically segmenting

both goals and missions that also allows for interleaved tasks, which they found to occur in 17% of tasks. Boldi et al. [2] describe the creation of query-flow graphs from query logs and show how they can be used to automatically identify chains of queries forming search tasks. Automatically detecting the set of queries that belong to the same task is a fundamental step for improving query suggestions or for a better choice of bidding terms for advertising. Radlinski and Joachims [11] consider tasks—or query chains—to aid a document ranking function.

3 Inferring User Intents

It is generally believed that inferring users' intents is difficult due to the fact that users do not express themselves clearly in the form of queries. In [5] against the general belief, the authors argue that users are capable of articulating their intents by queries. This claim was indeed confirmed by a preliminary study that reveals that in more than 78% of the cases users queries were demonstrative of their intents. The real intent of the user was inferred by the set of all the activities and interactions related to intent satisfaction. The authors propose a principled way to study the problem in the context of user goals [8, 2]. The terms goal and intent might be interchangeably used with the understanding that goals, comprised by a single query or multiple queries, are representative of atomic needs. The authors solve two different, though related, problems: understanding if the user was able to articulate her search goal by a query and identifying the query expressive of that intent. The two problems were formulated by a combination of behavioral, contextual and lexical features. The proposed models achieve 69% AUC on categorizing the multi-query goals and 62% AUC on single-query goals. Furthermore, the task of identifying the query that evinces the intent has a performance score of 81% AUC. These are very promising results given the highly challenging nature of the problem.

4 Supporting Complex Intents

As already stressed, users sometimes cannot see their needs immediately answered by search results, simply because these needs are too complex and involve multiple aspects that are not covered by a single web page and hence can not be expressed by a single query. Topics in domains such as education, travel or health, often require users browsing many different pages in order to accomplish the task they have in mind. Donato et al. [4] refer to this type of complex activities as "research missions". Research missions account for 10% of users' sessions and more than 25% of all query volume, as verified by a manual analysis that was conducted by Yahoo! editors. In [4] it was demonstrated that such missions can be automatically identified on-the-fly, as the user interacts with the search engine, through careful runtime analysis of query flows and query sessions. The on-the-fly automatic identification of research missions has been implemented in Search Pad, a Yahoo! application meant to help users keeping trace of results they have consulted. Its novelty however is that unlike previous notes taking

products, it is automatically triggered only when the system decides, with a fair level of confidence, that the user is undertaking a research mission and thus is in the right context for gathering notes. The analysis presented in [4] is one of the first example of session-awareness methodology in which user intent modeling is conducted by changing the level of granularity of the analysis, from an isolated query to a list of queries pertaining to the same research missions so as to better reflect a certain type of information needs.

5 Supporting Transactional Intents

Transactional queries are characterized by distinctive elements that differentiate them from navigational and informational ones. In [1], these distinctive elements were analyzed and used to develop a template-based methodology with the objective of directly supporting transactional queries and speed up tasks accomplishment. Such a methodology matches $n-$grams of lemmatized query terms against hierarchical dictionaries like WordNet and Wikipedia. Matched $n-$grams are hence substituted with the categories in order to generate a set of candidate "templates". The authors propose a probabilistic model to estimate the likelihood of each template to be generated by transactional queries and select the most likely ones to represent that transactional intent. Such a methodology can be seen as a first step in the attempt to change the current "informational" business model of web search engines. The main objective is to understand from the template the category to which the task belongs and to use the template to extract the information necessary to finalize the transaction. The query `tickets from NY to LA` clearly belongs to the `travel booking` category. All the queries that match the pattern `tickets from <city> to <city>` can be safely add to the same category. Such a pattern is responsible of deciding which application must be triggered for the booking process but, in order to finalize the transaction, the application needs to know two auxiliary inputs i.e. the origin (`from <city>`) and destination (`to <city>`). A comprehensive experimental study was conducted over eight different categories with a clear transactional intent varying from ticket booking and restaurant reservation to software or music download. The patterns were evaluated against a sample of queries randomly obtained from eight months of data extracted from Yahoo! query-logs. The results demonstrate that the methodology detects the transactional queries automatically and assigns them to the correct transactional category with a precision ranging from 0.7 to 0.98 depending on the category of interest.

6 Conclusions

In this short paper we presented some of the new lines of research conducted by the User Intent Analysis Group at Yahoo! Labs who has focused on understanding and modeling user intents. The common denominator for the most of the described models is a session-awareness methodology that has been changing the level of granularity of intent modeling, from an isolated query to a list of

queries pertaining to the same missions. This methodology is general and it is our strong belief that it is likely to play, in the near future, a fundamental role in many on-line tasks like detection of mission similarity or prediction of goal success and off-line task like partitioning users activity in topics or user behavior profiling.

References

1. A. Aashkan, P. Donmez, and D. Donato. Automatic rule extraction to identify transactional queries. *Submitted for publication*, 2011.
2. P. Boldi, F. Bonchi, C. Castillo, D. Donato, A. Gionis, and S. Vigna. The query-flow graph: model and applications. In *CIKM'08: Proceedings of the Information and Knowledge Management Conference*, pages 609–618, October 2008.
3. A. Broder. A taxonomy of web search. *SIGIR Forum*, 36(2):3–10, 2002.
4. D. Donato, F. Bonchi, T. Chi, and Y. S. Maarek. Do you want to take notes? Identifying research missions in Yahoo! search pad. In *WWW '10: Proceedings of the 19th International Conference on World Wide Web*, pages 321–330, 2010.
5. D. Donato, P. Donmez, B. Dumoulin, and H. Feild. Users are not lazy: Exploiting activity of articulate users to infer search intents. *Submitted for publication*, 2011.
6. M. R. Herrera, E. S. de Moura, M. Cristo, T. P. Silva, and A. S. da Silva. Exploring features for the automatic identification of user goals in web search. *Information Processing and Management*, 46(2):131–142, 2010.
7. B. J. Jansen, D. L. Booth, and A. Spink. Determining the informational, navigational, and transactional intent of web queries. *Information Processing and Management*, 44:1251–1266, May 2008.
8. R. Jones and K. L. Klinkner. Beyond the session timeout: automatic hierarchical segmentation of search topics in query logs. In *CIKM '08: Proceedings of the 17th ACM conference on Information and knowledge mining*, pages 699–708, New York, NY, USA, 2008. ACM.
9. I.-H. Kang. Transactional query identification in Web search. In *AIRS '05: Proceedings of Asian Information Retrieval Symposium*, 2005.
10. U. Lee, Z. Liu, and J. Cho. Automatic identification of user goals in web search. In *WWW '05: Proceedings of the 14th International Conference on World Wide Web*, pages 391–400, New York, NY, USA, 2005. ACM.
11. F. Radlinski and T. Joachims. Query chains: learning to rank from implicit feedback. In *KDD'05: Proceedings of the eleventh ACM SIGKDD International Conference on Knowledge discovery in data mining*, pages 239–248, New York, NY, USA, 2005. ACM Press.
12. D. E. Rose and D. Levinson. Understanding user goals in web search. In *WWW '04: Proceedings of the 13th International Conference on World Wide Web*, pages 13–19, New York, NY, USA, 2004. ACM.