The New Frontier of Web Search Technology: Seven Challenges

Ricardo Baeza-Yates[1], Andrei Z. Broder[2], and Yoelle Maarek[3]

[1] Yahoo! Research, Barcelona, Spain
[2] Yahoo! Research, Santa-Clara, CA
[3] Yahoo! Research Haifa, Israel

Abstract. The classic Web search experience, consisting of returning "ten blue links" in response to a short user query, is powered today by a mature technology where progress has become incremental and expensive. Furthermore, the "ten blue links" represent only a fractional part of the total Web search experience: today, what users expect and receive in response to a "web query" is a plethora of multi-media information extracted and synthesized from numerous sources on and off the Web. In consequence, we argue that the major technical challenges in Web search are now driven by the quest to satisfy the implicit and explicit needs of users, continuing a long evolutionary trend in commercial Web search engines going back more than fifteen years, moving from relevant document selection towards satisfactory task completion. We identify seven of these challenges and discuss them in some detail.

1 The Evolution of Commercial Search Engines

Commercial Web search engines have gone through a significant evolution in the last 16 years. We have identified three major stages in their evolution according to the type of data on which the technology has focused at the time:

- **On-Page Data:** First generation engines, like Excite, Lycos, or AltaVista in the middle of the 90's, used standard information retrieval models on crawled pages and used the *on-page* textual data almost exclusively, thus supporting only a syntactic match between queries and documents. The key paradigm was to parse HTML pages, possibly assigning more weights to important sections such as titles and abstracts, or author supplied keywords, and use usual *tf x idf* methods to compute relevance of pages within the Web corpus seen as a flat collection. The main challenge was scale and speed with relevance taking a second seat.
- **Web Graph Data:** Around 1998 a new approach, made popular by Google and eventually adopted by all engines, started to exploit *off-page* Web specific data. Three types of off-page data were leveraged: (1) link (or connectivity) data, initially simply in the form of in-degree (the number of links to a page), later through an analysis of the entire Web graph. The use of this data was based on the idea that "people vote with their links", [13, 4]; (2) anchor text data, that is, how people refer to a page on *other pages*, as a form of surrogated text abstraction of the target page. Here the idea was that "people vote with their labels"[10]; and finally (3) "click-through" data, that is, what results users clicked on as a form of implicit voting, the

S. Ceri and M. Brambilla (Eds.): Search Computing II, LNCS 6585, pp. 3–9, 2011.

idea being that "people vote with their clicks" [12], something that DirectHit used as early as 1997.

Simultaneously an infrastructure revolution took place based on distributed processing over a huge numbers of commodity computers. This novel infrastructure and its dedicated software enabled a "scale revolution" and allowed to crawl, index, store and serve more data than ever before, while respecting response time, latency and freshness constraints.

– **Usage Data:** Finally the third stage, which we are still experiencing, attempts to answer "the need" behind the query, that is, the unexpressed intent that drove the user to make a particular query in the first place. This phenomenon is expressed in multiple manners: First, the engine tries to guess what type of information best answers the intent and to this end, multiple sources of data are integrated in the result page, for example images, maps, videos, stock quotes, weather reports, current prices, tweets, news, etc. Second, additional query assistance tools, deployed both before and after the query is processed are becoming prevalent: for instance, query spell correction à la *did you mean* or query completion after the user entered only a few characters, are deployed on all major search engines. Similarly, results exploration tools (*e.g.*, narrowing search results by type, source, date, translation, etc.) are becoming common place. This is achieved by focusing on yet another type of data: usage data. While previously usage data had been aggregated at page level, the attention has now moved from individual pages to individual users. Users' behavior at every stage of their interaction with the search engine, and even their post-interaction behavior (via browser toolbars, beacons, etc.) can be captured in very detailed logs for an enormous number of queries. By studying users' behavior *after* they make a particular query (*e.g.*, query reformulations, clicks, browsing time), one gains an understanding of users' intent. In practice, efficient statistical methods are and can be used to create adequate pre- and post-search assistance tools.

By monitoring users' activities and gathering usage data at a very large scale, search engines serve users at two levels: *implicitly* by trying to guess unexpressed intent, as if the engine was reading the user's mind and *explicitly* by offering interactive tools, which not only make the search experience more attractive but also provide additional hints regarding the user's intent. We believe these two directions represent the new frontier of Web search and are associated with a number of technical challenges. We list below a few of these challenges and areas where progress is being made and more innovation is to be expected. For readers interested in a detailed coverage of Web retrieval we suggest the chapter by Baeza-Yates and Maarek in [3].

2 Ongoing and New Challenges

As discussed earlier, monitoring users' activities on a very large scale allows to better answer implicit and explicit information needs [5, 19] and more specifically query intent. We have identified seven challenges that we believe represent opportunities for further research and innovation, some already seeing incremental progress happening on a regular basis and others demanding drastic departure from previous art. We have ranked them below by their order of (possibly future) appearance in the Web.

2.1 Query Assistance

Query assistance tools first appeared on the search engine results page, offering alternate query forms in case the user was not satisfied with the returned results. These alternate queries took two possible forms: first related queries, typically derived from the results themselves, and soon after query spelling suggestion that leverages usage data [6]. The novelty in the now famous "did you mean" feature for instance consisted in its learning from usage data rather than using a fixed dictionary. Multiple techniques are used today such as counting most frequent queries at a small edit distance of the original query[1] or looking at query reformulation as users tend to correct themselves if their original query was misspelled [16]. Leveraging usage data via query-log analysis at a large scale gave and is still giving excellent results. It is also one of the best examples of the "wisdom of crowds" [21]. It took a longer time however for the attention to move to the core search box and try helping users formulate their queries even before the query is issued. The first major query assistance tool appeared in 2004 when Google offered " Suggest"[2], as an experimental feature on Google Labs. Thanks to the scale revolution, it was suddenly possible to use as completion dictionary a large query log, which gave the impression of an uncontrolled vocabulary experience. At that stage the corpus was static, but progress in infrastructure allowed Yahoo! Search to launch "Search Assist" in 2007 and the following year Google to launch Suggest on google.com and youtube.com. Nowadays, most engines offer this feature and keep improving it, with better freshness, coverage, locality, instant previews, etc. This represented a critical stage in helping users express their intent, based on the conjecture that the odds are good that a previous user with similar intent found a good query to express it.

2.2 Contextualization

Another trend in recent years has been the *contextualization* of the answer. We use the term contextualization in a generic manner to cover (1) localization (geographical and/or language contextualization), (2) personalization (user contextualization), (3) socialization (that is, take in account the social context), and (4) query intention (intent contextualization), among others. Considering that most users interact little and do not explicitly authorize individual identification (by signing-in for instance), personalizing is a difficult task, which impacts only a small percentage of users. In contrast, intent contextualization relies on analyzing the usage data originating from users conducting the same task. It is both more pragmatic (as it does not require signing-in or explicit authorization), and more effective as it can applied to larger populations of users and thus some limited form of the previous mentioned "wisdom of crowds" intuitions can be applied. Most of all, privacy infringement risks are significantly reduced as no single user is isolated, and techniques are applied to groups of people (small crowds). The effectiveness of these methods come from the fact that while users are all different in their heterogeneous needs or facets, on each facet they are not that different from other users

[1] A well known Google example shows that the correct spelling for the query "Britney Spears" is more frequent by an order of magnitude than its immediate follower, see http://www.google.com/jobs/britney.html

[2] Now called autocomplete [22].

and perform similar tasks, their uniqueness comes from the combination of these facets and on when, how long and how well they conduct those tasks. The challenge here is then to better detect query intent and better contextualize intention. Contextualizing the results affects search results display and the overall user experience (*e.g.*, geographical contextualization may require displaying an interactive map within the search results page) and hence triggers the next challenge: how to present different types of results. Regarding the social aspect, for some search needs the social context is clearly relevant as the Web is a communication media that is owned in a large extent by its users through the Web 2.0 [18].

2.3 Universal Search

Another significant step in guessing the intent of the query is not to require from the user to specify what type (*e.g.*, image, video, map, etc.) or source of data (*e.g.*, news, blogs, encyclopedia, etc.) s/he is more interested in but simply guess for a given query what types and sources should be shown. The goal here is to integrate rich and complex data source in a semi-transparent manner. This concept was coined "universal search" by Marissa Mayer in [14] and continues seeing a great deal of progress. It presents multiple challenges, as it requires "comparing apples and oranges", and more specifically deciding what sources should be probed, how many results from relevant sources should be shown, where in the ranked list these results should be slotted if at all, etc. This area becomes even more intriguing with real-time feeds such as tweets for which relevance needs to be estimated in almost real-time. One open problem is the screen layout for the different types of results if you want to move away from the classical sequential list of ten results. This research area is now called "aggregated search" and naturally leads to the next challenge.

2.4 Web of Objects

A more recent challenge consists of departing from the usual result triplet (title,snippet, link) as a surrogate of a given Web page and returning instead the object that really satisfied her needs. A typical example when searching for an artist is to get as a result not an heterogeneous list of links to his official site, images, videos, fan club, wikipedia entry, lyrics pages, etc. but rather a composite object that integrates all the possible facets that should be relevant to the user. The same goes for famous athletes (for whom the user would like to get team information, recent stats, photos, etc.), restaurants (address, map, opening hours, reviews,etc.), travel destinations (slideshows, weather, hotels, places to see, etc.). The key idea behind the concept of Web of Objects is that the individual pages that typically form the Web are exploded into individual objects that can be recomposed into a synthetic page, which is then shown to users as a search result. Thus, in our previous athlete example, a user searching today for the Viking quarterback "Brett Favre" will see on a Yahoo! Search, as top result, a synthetic concise "mini page" consisting of various objects such as 2010-2011 stats displayed as a dashboard (with QB rating, number of touch downs etc.), extracted from a given Web site), a profile generated from another Web site, links to news, games logs, Scores and Schedules, etc., all extracted from various sites as needed. More generally, integrating Web derived knowledge, well

beyond entity extraction, towards building and representing interrelationships between known entities, will enable users to search not only the"Web of Pages" but the "Web of Objects" as detailed in [2, 7].

2.5 Post-Search Experience

As we focus on intent, we must address needs that go beyond simple information discovery. We are still doing very little with results so far. We can "star" them [9], translate the associated page via a single link on Google, or share video results in a social network or via messenger or email in Bing. In addition, major search engines now offer the ability to narrow search results according to various facets [15]. Some results can be displayed in various microformats [17] using Yahoo! SearchMonkey [20] or Google's rich snippets [11] via common agreement between publishers and search engines. Yahoo! Search Pad [8] goes one step further as it automatically gathers clicked results of a same search session and allow users to annotate/edit/share such pads. Yet, it seems that there is a great deal of opportunities for engines to offer additional tools that would facilitate post-search experience, mostly for better exploration and manipulation (filtering, extracting, etc.) of results so as to better satisfy the underlying query intent. This challenge is also related to "universal search".

2.6 Application Integration

A natural extension of the post-search experience would be to go further with result manipulation and facilitate the integration of third party applications to enable a richer, more diversified, and more satisfying user experience. Underlying intent might involve a series of tasks, such as planning for a holiday, organizing a birthday celebration, etc. where search results represent only raw data that need to be digested and processed to satisfy the intent. This mostly unexplored area presents fascinating technical challenges. A small but significant step in this direction was conducted by Yahoo! Search recently via its "QuickApps" mechanism. Third-party applications are offered on the left trail of the results page by partners such as Netflix, the popular US-based flat-rate movie rental/online video streaming services, or OpenTable the online restaurant reservation service. Such applications can be triggered for specific queries and pre-populated with the needed parameters, once a " movie intent" or "restaurant intent" are identified. This application integration approach focuses on facilitating the task behind the query in order to better satisfy users.

2.7 Implicit Search

Finally, the most intriguing of all challenges is "implicit search", which aims at addressing users' needs without requiring them to express a query. As search engines better and better understand their users by monitoring their activities and building sophisticated users models, and multiple contextual signals are accessible (via sensors, GPS, cell information, etc.) one can envision scenarios in which the need can be identified by a simple click or simply as a side effect of another action. A taste for implicit search is offered today in book selling, like Amazon, or movie rental, such as NetFlix, services

with their recommendation services. Another example is the recently launched Priority Inbox in Gmail [1]. These are only preliminary steps in this direction. We believe that with the fast penetration of smart phones and cloud computing where all devices can be associated with a single user, search engines will have at their disposal a plurality of signals that will make the difference in personalization and contextualization. One can envision implicit search mechanisms being triggered within various applications as mentioned above, or related content being pushed to users in appropriate contexts. However, most wild scenarios will remain unrealistic if privacy concerns are not addressed and answered in a satisfying manner.

3 Conclusion

Web search has now become a mature technology with high penetration to the general population in most developed countries. We believe that the focus of innovation should move towards the "before" and "after" stages in web search with endless possibilities. Numerous challenges are involved, including but definitely not limited to the seven we have listed here, with multiple research and technology development opportunities. Overall, all of them are focused in improving the overall search experience of the user. Nevertheless, some of these new results will also improve user experience in general.

References

[1] Aberdeen, D.: Email overload? try priority inbox. The Official Gmail Blog (August 2010), http://gmailblog.blogspot.com/2010/08/email-overload-try-priority-inbox.html

[2] Baeza-Yates, R., Raghavan, P.: Chapter 2: Next generation web search. In: Ceri, S., Brambilla, M. (eds.) Search Computing. LNCS, vol. 5950, pp. 11–23. Springer, Heidelberg (2010)

[3] Baeza-Yates, R., Ribeiro-Neto, B.: Modern Information Retrieval. Addison-Wesley, Harlow (2010)

[4] Brin, S., Page, L.: The anatomy of a large-scale hypertextual web search engine. In: Proceedings of the 7th International Conference on World Wide Web, Brisbane, Australia (1998)

[5] Broder, A.: A taxonomy of web search. SIGIR Forum 36(2) (Fall 2002)

[6] Cucerzan, S., Brill, E.: Spelling correction as an iterative process that exploits the collective knowledge of web users. In: Proceedings of Empirical Methods in Natural Language Processing, Barcelona, Spain (July 2004)

[7] Dalvi, N.N., Kumar, R., Pang, B., Ramakrishnan, R., Tomkins, A., Bohannon, P., Keerthi, S., Merugu, S.: A web of concepts. In: PODS, pp. 1–12 (2009)

[8] Donato, D., Bonchi, F., Chi, T., Maarek, Y.: Do you want to take notes? identifying research missions in yahoo! search pad. In: Proceedings of the 19th International Conference on World Wide Web, Raleigh, North Carolina, USA, pp. 321–330 (2010)

[9] Dupont, C.: Stars make search more personal. The Official Google Blog, March 3 (2010), http://googleblog.blogspot.com/2010/03/stars-make-search-more-personal.html

[10] Eiron, N., McCurley, K.S.: Analysis of anchor text for web search. In: Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Informaion Retrieval, SIGIR 2003, pp. 459–460. ACM, New York (2003)

[11] Goel, K., Guha, R., Hansson, O.: Introducing rich snippets. Google Webmaster Central Blog (May 2009), http://googlewebmastercentral.blogspot.com/2009/05/introducing-rich-snippets.html
[12] Joachims, T.: Optimizing search engines using clickthrough data. In: Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD 2002, pp. 133–142. ACM, New York (2002)
[13] Kleinberg, J.: Authoritative sources in a hyperlinked environment. In: Proceedings of the ACM-SIAM Symposium on Discrete Algorithms, pp. 668–677 (1998)
[14] Mayer, M.: Universal search: The best answer is still the best answer.The Official Google Blog (May 2007), http://googleblog.blogspot.com/2007/05/universal-search-best-answer-is-still.html
[15] Mayer, M., Menzel, J.: More search options and other updates from our searchology event. The Official Google Blog (May 2009), http://googleblog.blogspot.com/2009/05/more-search-options-and-other-updates.html
[16] Merrill, D.: http://www.youtube.com/watch?v=syKY8CrHkck#t=22m11s at timestamp 22m11s
[17] Mika, P.: Microsearch: An interface for semantic search. In: Proceedings of the SemSearch 2008 Workshop on Semantic Search at the 5th European Semantic Web Conference, Tenerife, Spain (June 2008), http://sunsite.informatik.rwth-aachen.de/Publications/CEUR-WS/Vol-334/
[18] Ramakrishnan, R., Tomkins, A.: Toward a peopleweb. IEEE Computer 40(8), 63–72 (2007)
[19] Rose, D.E., Levinson, D.: Understanding user goals in web search. In: WWW 2004: Proceedings of the 13th International Conference on World Wide Web, pp. 13–19. ACM, New York (2004)
[20] Searchmonkey, http://developer.yahoo.com/searchmonkey/
[21] Surowiecki, J.: The Wisdom of Crowds: Why the Many Are Smarter Than the Few and How Collective Wisdom Shapes Business, Economies, Societies and Nations. Random House (2004)
[22] Wright, J.: This week in search 10/16/10: Renaming google suggest. The Official Google Blog (October 16, 2010)