SEEKS
Project Manifesto

Emmanuel Benazera

Sylvio Drouin

Version 1.0

October 10th, 2006

Scope: This document exposes the main rationale behind the Seeks project, an open-source pattern matching peer-to-peer overlay network for social web search. The authors hope is that by A) exposing their views and concerns to the community, about the current state of web search technologies in relation to the accuracy, privacy and origin of information; B) introducing novel web search models and algorithms; and C) presenting new solutions to the problem of the corporation's unfair advantage over individual users when deploying content on the world wide web; the free software movement will play a crucial role in the future of web search through models and principles such as those defended by the Seeks project. In all cases, we do understand that the software industry only develops certain projects, those that lead to profit, mainly financial, leaving several great ideas to oblivion. Bringing free software and democracy to web search faces many challenges, ranging from the protection and open licensing of data (web indexes, queries), to the public control of these indexes, and the open sourcing of the software. As such, the Seeks project is a response to the software issue. While its architecture is designed to protect the users' rights, in the end, we hope that the project itself helps focusing the community's attention on solutions to the proper licensing of web indexes and user queries.

I - The Internet and the end-to-end architecture

The Internet relies on an end-to-end architecture, where the power lies in the leaves of the network, such as web servers or user machines. Over the past decade, we observed the rise of two new major Internet topologies, each having contributed to bring the web sphere to its current state: first, the gateway-like topology where servers gather the network traffic and redistribute it to the leaves, namely the search engines; second, the bag-like topology where the traffic gets trapped within a single set of servers, namely the so-called social web-communities. We would not be concerned with these topological trends if they did not appear to us as both risky and inefficient in the long term. The most prevalent problem of course lying in the massive personal and public information collections now being held by the businesses initially responsible for the above mentioned topological changes.

Whenever we're asking a question to somebody, whoever that person can be, there are good chances we are revealing something about ourselves in the process: our interests, our opinions, etc. It is how the chain of trust is established. This chain of trust should also exist when we're querying a search engine or accessing our webmail, but in this case the interlocutor is not another human being but a set of sophisticated and confidential algorithms that will record, re-use and most probably distort all the information accumulated about us.

We believe the Web has reached a point where the chain of trust has broken down. We believe that the process by which centralized entities, with high-traffic capacity, obtain private profile and behavioral data, has become transparent to the point where users are lured into using free, and most of the time essential services; gradually come to rely on them; in the end to be coerced (through clever web authoring) into revealing extensive personal information without ever realizing they're doing so. This is the reason why users need to invent ways to protect themselves, share information, and evaluate this information based on the collective trust of all users rather than based on the results of few corporations greed-influenced algorithms. For that reason, **we believe that user queries to search engines should be shared with the global search community rather than being recorded by powerful corporate entities**. In the Seeks project this is easily done by sharing the search queries over a distributed hash table (a DHT, also known as a peer-to-peer overlay network). We understand the difficulties and opposition users may (and certainly will) have to the sharing of queries, but we also believe that the benefits far outweighs the risks.

II - Search engines and their problems

We believe that search engines power has reached a certain plateau: first, it is very common knowledge and experience that even new algorithms, while returning satisfying relevant results to simple queries (finding a shop, a band, etc...), perform poorly on content that is less well tailored to direct exposure to the engine crawlers (forum discussions, blog comments, ...). Also, the same algorithms are known to be easily fooled (e.g. by building fake pages). Second, given the rising complexity of the ad hoc rules of website ranking and elimination by the search engines, we observe that branding, advertising oriented and other carefully tailored websites, force commercial contents and their servers to the front row by buying the expensive services of web publishing companies, while leaving in the dark the mass of user generated information. Drawing from these remarks, we believe that the web has been partly hijacked from the end users, and that there should be a serious attempt by the free software community to return it to the general population. We believe this should start with the web searching and publishing experiences. While our views may be perceived as naive and rebellious by some, we believe that the Seeks project, in creating both a framework for initial discussions and an open, transparent platform for the integration of social search technologies, will eventually be accepted as a necessary step toward the creation of a fairer and more social web experience.

As such, the Seeks project proposes to share the search queries among users, naturally building a

collaborative social filter on top of the main search engines and their results. **Today, the lack of social sharing leads to masses of users doing the same searches over and over again, all over the world, while remaining alone.** We believe that bringing them together to share their experiences should lead to an easier and better convergence between search queries and web content, and hopefully a more enjoyable experience of the world wide web.

In other words, we understand the need for a measure of the fit between queries and results, and that such a measure should rely on user ratings rather than on automated procedures. The tremendous amount of junk or inadequate results can be mitigated by a collective effort. We do believe so, because through experience and past projects, including state of the art AI and interface design for the most advanced public and private companies and labs, we came to understand that algorithms cannot, and will not, lead to a satisfactory handling of human generated data, more especially in dynamic, ever changing, environments. This should not be perceived as an acceptance of failure, but as the beginning of a long-term effort to provide individual users (not just corporations and advertising agencies), with access to the tools, technologies and sophisticated algorithms that are required to be again in control of the surfing experience and in doing so, increase one's influence on how and where information flows on the world wide web.

III - Three steps for getting more out of the web and its content.

As of today, there is a three steps road map for the Seeks project. We believe each stage introduces essential functionalities, but we leave doors open for users to define new ones.

First, Seeks provides the basic collaborative functionalities on top of existing search engines by connecting people that search the web with similar queries. The similarity-based pattern-matching technique used is known as locality sensitive hashing (LSH) and we distribute it among the peers. Existing search engines results are re-sorted and enhanced with the information fetched from the peer-to-peer network, such as ratings, other relevant results, related queries of interest, and direct chat opportunity is provided to users performing the same, or approached queries. Enhanced with a clean and modern interface within the browser, we believe these core capabilities should be of interest to number of people, and we wish they draw in users, testers and developers in sufficient numbers.

The second step is key to the Seeks project, and introduces what we believe are its most beneficial features. Seeks will propose a self-publishing mechanism accessible to anybody with a browser and an Internet connection. Instead of relying on a search engine for linking keywords to web contents

(through crawling and indexing), Seeks will let the users register any URL using their own set of keywords (in other words, their own queries). Users querying the peer-to-peer network (DHT) will thus be recommended web content *without* using any existing search engine. Technically, this operation is the combination of a DHT lookup plus a selection on the peer of interest, and should not cost much more than a file lookup from your favorite peer-to-peer front end, (i.e. pretty fast). Users that would register their personal web content or that of others under unsatisfactory keywords or queries would see their keyword associations naturally rated down by other users, in a move that we believe should lead to a better match of keywords and queries to the true content of a web page. Finally, and nonetheless, we're in the process of defining the setup of virtual marketplaces over keywords for publishing web content, at each of the DHT peers. These marketplaces would not rely on money but on fame instead, understood as a measure of a user attachment to truth. Thus any hot content recommended to Seeks users would come from a fair collective pre-selection among bidding users.

The third and final step proposes a decentralized web information index to gradually re-capture public information currently stored in private corporate facilities. We propose to implement small software extensions to common WEB servers, such as Apache. These extensions would allow web servers to locally index their webpages and share the indexes with other web servers and users, in a decentralized manner, on the Seeks network. The consequence is that over time, Seeks would evolve a parallel search engine, processing queries against a decentralized database of information rated by the community for the community.

This manifest is a short exposition of the rational and aims of the Seeks project. By exposing them we expect interested readers to discuss the forces and drawbacks of the project among themselves and with us, and if they feel it is worth their time, to get in touch with us and help us defining, redefining, and developing all above features.

Emmanuel Benazera & Sylvio Drouin