The Anatomy of a Large-Scale *Social* Search Engine

Damon Horowitz
Aardvark
damon@aardvarkteam.com

Sepandar D. Kamvar
Stanford University
sdkamvar@stanford.edu

ABSTRACT

We present Aardvark, a social search engine. With Aardvark, users ask a question, either by instant message, email, web input, text message, or voice. Aardvark then routes the question to the person in the user's extended social network most likely to be able to answer that question. As compared to a traditional web search engine, where the challenge lies in finding the right *document* to satisfy a user's information need, the challenge in a social search engine like Aardvark lies in finding the right *person* to satisfy a user's information need. Further, while trust in a traditional search engine is based on authority, in a social search engine like Aardvark, trust is based on intimacy. We describe how these considerations inform the architecture, algorithms, and user interface of Aardvark, and how they are reflected in the behavior of Aardvark users.

1. INTRODUCTION

1.1 The Library and the Village

Traditionally, the basic paradigm in information retrieval has been the library. Indeed, the field of IR has roots in the library sciences, and Google itself came out of the Stanford Digital Library project [18]. While this paradigm has clearly worked well in several contexts, it ignores another age-old model for knowledge acquisition, which we shall call "the village paradigm". In a village, knowledge dissemination is achieved socially — information is passed from person to person, and the retrieval task consists of finding the right person, rather than the right document, to answer your question.

The differences how people find information in a library versus a village suggest some useful principles for designing a social search engine. In a library, people use keywords to search, the knowledge base is created by a small number of content publishers before the questions are asked, and trust is based on authority. In a village, by contrast, people use natural language to ask questions, answers are generated in real-time by anyone in the community, and trust is based on intimacy. These properties have cascading effects — for example, real-time responses from socially proximal responders tend to elicit (and work well for) highly contextualized and subjective queries. For example, the query "Do you have any good babysitter recommendations in Palo Alto for my 6-year-old twins? I'm looking for somebody that won't let them watch TV." is better answered by a friend than the library. These differences in information retrieval paradigm require that a social search engine have very different architecture, algorithms, and user interfaces than a search engine based on the library paradigm.

The fact that the library and the village paradigms of knowledge acquisition complement one another nicely in the offline world suggests a broad opportunity on the web for social information retrieval.

WWW2010, April 26-30, 2010, Raleigh, North Carolina.

1.2 Aardvark

In this paper, we present Aardvark, a social search engine based on the village paradigm. We describe in detail the architecture, ranking algorithms, and user interfaces in Aardvark, and the design considerations that motivated them. We believe this to be useful to the research community for two reasons. First, the argument made in the original Anatomy paper [4] still holds true — since most search engine development is done in industry rather than academia, the research literature describing end-to-end search engine architecture is sparse. Second, the shift in paradigm opens up a number of interesting research questions in information retrieval, for example around expertise classification, implicit network construction, and conversation design.

Following the architecture description, we present a statistical analysis of usage patterns in Aardvark. We find that, as compared to traditional search, Aardvark queries tend to be long, highly contextualized and subjective — in short, they tend to be the types of queries that are not well-serviced by traditional search engines. We also find that the vast majority of questions get answered promptly and satisfactorily, and that users are surprisingly active, both in asking and answering.

Finally, we present example results from the current Aardvark system, and a comparative evaluation experiment. What we find is that Aardvark performs very well on queries that deal with opinion, advice, experience, or recommendations, while traditional corpus-based search engines remain a good choice for queries that are factual or navigational.

2. OVERVIEW

2.1 Main Components

The main components of Aardvark are:

1. *Crawler and Indexer.* To find and label resources that contain information — in this case, users, not documents (Sections 3.2 and 3.3).