Social Search

Introduction to Information Retrieval
INF 141/ CS 121
Donald J. Patterson
The Anatomy of a Large-Scale Social Search Engine

Damon Horowitz
Aardvark
Jaron@eurodigitalant.com

Sepandar D. Kamvar
Stanford University
seckamvar@stanford.edu

ABSTRACT

We present Aardvark, a social search engine. With Aardvark, users ask a question, either in instant message, email, web log, text message, or voice. Aardvark then queries the question to the person in the user’s extended social network most likely to be able to answer that question. Aardvark can be seen as a traditional web search engine, where the challenge lies in finding the right answer to a query. Aardvark is a social search engine, where the challenge lies in finding the right person to ask the query. Aardvark is based on a new paradigm, a social network search engine like Aardvark, that is based on social networks. We describe how these constellations of knowledge and reputation enable users to search in the offline world without sacrificing speed and access to social information.

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 science and Google just came out of the Stanford Digital Library project (LD). While this paradigm has surely worked well in several settings, it ignores and maybe even exploits knowledge networks, which we do not have in the village paradigm. In a village, knowledge dissemination is social, not relational; information is passed from person to person, and the topics can be topics of finding the right person, rather than the right document, to answer your questions.

The differences between finding information in a library versus a village suggest some useful information for designing a social search engine. In a library, people seek information in a structured environment. In a village, people seek information in a networked environment. The knowledge is based on authority, and trust is based on authority. In a village, people seek information in a natural language, and the questions are natural and, most importantly, meaningful. These properties have an overwhelmingly important effect on searching, especially searching with social networks. For example, people are more likely to ask questions in social networks, and the questions are likely to be meaningful. For example, the query “Let’s have good ideas for my 6-year-old today” is likely to be asked by someone who isn’t just searching for information. These differences in information retrieval paradigm suggest that social search engines have different architectures, algorithms, and user interfaces than a search engine based on the library paradigm.

The fact that the library and the village paradigms of knowledge and reputation can be used in the offline world suggests a broad opportunity on the web for social information retrieval.

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 algorithm, and user interface of Aardvark. We also describe the design considerations that motivated them.

We believe this to be useful to the researchers community for two reasons. First, the approach taken in the original Aardvark paper (LD) will have input into what social search engines development is like in industry today, since academia is the standard literature source. Second, the skills in designing complex social systems, for example around reputation classification, implicit network construction, and conversation design is

2. OVERVIEW

2.1 Main Components

The main components of Aardvark are:

1. Queries and Indicators: To find and label resources that contain information — it will use, users, and documents (Sections 2.2 and 2.3).
Aardvark

- Web IR
  - Input is a query of keywords
  - Search is over documents
  - Trust is based on authority
  - Mental model is a library
Aardvark

• Web IR
  • Input is a query of **keywords**
  • Search is over **documents**
  • Trust is based on **authority**
  • Mental model is a **library**

• Social Search
  • Input is a **question**
  • Search is over **people**
  • Trust is based on **intimacy**
  • Mental model is a **village**
Aardvark

- Web IR
  - facts
  - navigation
  - transactions

- Social Search
  - opinion
  - advice
  - experience
  - recommendations
Components

- Crawler Indexer
- Query Analyzer
- Ranking Function
- UI
Aardvark

- Crawler/Indexer
- Users not documents
- Query Analyzer
- Understand the information need
- Ranking Function
- Pick the best resources
- UI
- To manage the conversation
Welcome to Aardvark - Sign on process

• After confirming a new user’s account
• A Social Graph is built
  • Facebook/LinkedIn connections
  • webmail connections
  • manual email invites
  • “group” aware
    • This is a work colleague, college friend, etc.
Welcome to Aardvark - Sign on process

- A knowledge bank is built on
  - self-identified expertise
  - friend identified expertise
  - home page identified expertise
  - facebook status update analysis
  - twitter status update analysis
  - observed Aardvark usage

- knowledge bank’s inverted index maps topic -> user
Aardvark

The Question

Ask a question and I'll find someone to answer

Are there any frameworks that use cloud computing that are similar in spirit to MapReduce? MapReduce enables a certain kind of computation in cloud computing. Are there other frameworks that are worth knowing about as well?

Send

Example questions

Ask a question and I'll find someone to answer

What's a good idea for a 13 year old's birthday party in Orange County for 10 kids that won't require me to mortgage my house?
The Question

- The question is acquired by various input channels:
  - text, web, IM, mobile, etc.
- The question is screened for obscenity
- The question is topic analyzed
  - topic is presented to asker for confirmation
- The question is passed to a routing engine
  - which ranks potential answers based on
  - social graph and expertise
Aardvark

Routing Engine

- Pick the best answerer

\[ p(u_i | q) = \sum_{t \in T} p(u_i | t)p(t | q) \]

- What is the probability that user i will answer question q?
- Marginalize over the topics in the question and the topic expertise of the user

- Pick the best pair of users

\[ p(u_i | u_j) \]

- Which user i is the most likely to give a good answer to user j?
Aardvark

Routing Engine

\[ s(u_i, u_j, q) = p(u_i | u_j) \cdot p(u_i | q) = p(u_i | u_j) \sum_{t \in T} p(u_i | t)p(t | q) \]

- Ranking function combines the two
- For a given query, q, by user j
  - what is the best user i to ask?
- Biases intimacy over authority
- Notice there is nothing like PageRank here
- The only real-time component is \( p(t | q) \)
Indexing People

- Figuring out $p(t|u_i)$
- Figuring out $p(u_i|u_j)$
Indexing People

- Figuring out $p(t|u_i)$
- Users self-identify topics they are “experts” in
- Others identify topics you are “experts” in
- Topics are mined from
  - Facebook
  - Twitter
  - Homepages
  - Blogs
Indexing People

- Figuring out $p(t|u_i)$
  - In unstructured text
    - an SVM classifies the general topic
    - an ad-hoc entity parser figures out specific topics
  - scaled by TF-IDF score
- topics are also mined from aardvark conversations
Indexing People

- The mined information is not for answering questions
- It’s for identifying people who can answer questions
- Topics are enhanced with social network information

\[ s(t|u_i) = p(t|u_i) + \gamma \sum_{u \in U} p(t|u) \]
Indexing People

- All of the scores for topics given a user are normalized
  \[ \sum_{t \in T} p(t | u_i) = 1 \]
- Bayes Law is used to invert the probability
  \[ p(u_i | t) = \frac{p(t | u_i) p(u_i)}{p(t)} \]
Indexing People

- Probability that $u_i$ will respond to $u_j$ 
  \[ p(u_i|u_j) \]
- Social Connections
- Demographic similarity
- Profile similarity
- Vocabulary similarity
- Chattiness similarity
- Verbosity similarity
- Politeness similarity
- Speed match
Pulling topics out of a question $p(t|q)$

- Real-time response needed
- Doesn’t have to be perfect, people pick up the slack
- Screening
  - Is it a question? (No -> reject)
  - Is it trivial? (Yes -> answer it)
  - Is it location sensitive?
    - Location is treated differently than topic
Pulling topics out of a question

- topics
  - Keyword to Topic Matcher
  - Taxonomy Topic Mapper
    - SVM classifier
  - Phrase to Topic Matcher
  - User Tag to Topic Mapper
    - If the user tags the question

\[ p(t|q) \]
Aardvark

User Interface

```
aardvark: You there? I have a question about **tennis** that I think you might be able to answer. (Type 'sure', 'pass', or 'busy'.)

me: sure

aardvark: (From Michael B./29/M/NewYork,US)
Does anyone know which health clubs in Manhattan have tennis courts? Or are all the good options in the other boroughs?
(Type 'pass' to skip, or 'more' for extra options.)

me: why

aardvark: I thought of you for Michael's question because
- Michael is a friend of your friend Sally Johnson.
- You know a lot about **racquet sports** and **fitness**
- You and he have similar interests in **sports**
(Type 'more' or change settings at http://vark.com/a/XXXXX)

me: Well there is always the Midtown Tennis Club on 8th ave @27th
if you really want to stay in manhattan -- but the quality isn't great.
You'd do just as well to use the public courts in Central Park. Or
another good option is to join NYHRC or NYSC in manhattan, and
use their courts in other boroughs...

aardvark: Great -- I've sent that to Michael. Thanks for the fast
answer! (Type 'Michael:.' followed by a message to add something, or
'more' for options.)
```
Aardvark

User Interface

![Image of Aardvark user interface]

Aardvark is a platform that allows users to ask and answer questions. The interface is designed to be user-friendly, with options to select different categories such as consumer electronics, cafes, and fishing. Users can ask questions and see open questions about various topics. The example question in the image is: "What's a cafe in SOMA with great coffee and food and wifi where I'd be able to sit down and get some work done?" The answer is provided by Angela Smith, aged 28, who is a female friend on Facebook.
<table>
<thead>
<tr>
<th>Question</th>
<th>Question</th>
</tr>
</thead>
<tbody>
<tr>
<td>&quot;What fun bars downtown have outdoor seating?&quot;</td>
<td>&quot;I've started running at least 4 days each week, but I'm starting to get some knee and ankle pain. Any ideas about how to address this short of running less?&quot;</td>
</tr>
<tr>
<td>&quot;I'm just getting into photography. Any suggestions for a digital camera that would be easy enough for me to use as a beginner, but I'll want to keep using for a while?&quot;</td>
<td>&quot;I need a good prank to play on my supervisor. She has a good sense of humor, but is overly professional. Any ideas?&quot;</td>
</tr>
<tr>
<td>&quot;I'm going to Berlin for two weeks and would like to take some day trips to places that aren't too touristy. Where should I go?&quot;</td>
<td>&quot;I just moved and have the perfect spot for a plant in my living room. It gets a lot of light from the north and south, but I know I won't be too reliable with watering. Any suggestions for plants that won't die?&quot;</td>
</tr>
<tr>
<td>&quot;My friend's in town and wants to see live music. We both love bands like the Counting Crows. Any recommendations for shows (of any size) to check out?&quot;</td>
<td>&quot;Should I wear brown or black shoes with a light brown suit?&quot;</td>
</tr>
<tr>
<td>&quot;Is there any way to recover an unsaved Excel file that was closed manually on a Mac?&quot;</td>
<td>&quot;I need to analyze a Spanish poem for class. What are some interesting Spanish poems that aren't too difficult to translate?&quot;</td>
</tr>
<tr>
<td>&quot;I'm putting together a focus group to talk about my brand new website. Any tips on making it as effective as possible?&quot;</td>
<td>&quot;I always drive by men selling strawberries on Stanford Ave. How much do they charge per flat?&quot;</td>
</tr>
<tr>
<td>&quot;I'm making cookies but ran out of baking powder. Is there anything I can substitute?&quot;</td>
<td>&quot;My girlfriend's ex bought her lots of expensive presents on anniversaries. I'm pretty broke, but want to show her that I care. Any ideas for things I could do that are not too cliche?&quot;</td>
</tr>
</tbody>
</table>
(Question from Brian T./22/M/Castro,SF) What is a good place to take a spunky, off-the-cuff, social, and pretty girl for a nontraditional, fun, memorable dinner date in San Francisco?

(+4 minutes -- Answer from Dan G./M/SanFrancisco,CA)  
Start with drinks at NocNoc (cheap, beer/wine only) and then dinner at RNM (expensive, across the street).

(Reply from Brian to Dan) Thanks!

(+6 minutes -- Answer from Anthony D./M/Sunnyvale,CA -- you are both in the Google group)  
Take her to the ROTL production of Tommy, in the Mission. Best show i've seen all year!

(Reply from Brian to Anthony) Tommy as in the Who's rock opera? COOL!

(+10 minutes -- Answer from Bob F./M/Mission,SF -- you are connected through Mathias' friend Samantha S.) Cool question. Spork is usually my top choice for a first date, because in addition to having great food and good really friendly service, it has an atmosphere that's perfectly in between casual and romantic. It's a quirky place, interesting funny menu, but not exactly non-traditional in the sense that you're not eating while suspended from the ceiling or anything
Empirical Results

- websites & internet apps
- music, movies, TV, books
- technology & programming
- local services
- product reviews & help
- restaurants & bars
- travel
- finance & investing
- home & cooking
- sports & recreation
- business research
- Aardvark
Empirical Results

- 0-3 min: 2.5 x 10^4
- 3-6 min: 2.0
- 6-12 min: 1.5
- 12-30 min: 1.0
- 30 min-1 hr: 0.5
- 1-4 hr: 0
- 4+ hr: 0
Empirical Results
Aardvark

Empirical Results

![Bar chart showing empirical results for different numbers of topics. The x-axis represents the number of topics (>2, >4, >8, >16, >32), and the y-axis represents the percentage (0%, 20%, 40%, 60%, 80%, 100%). The chart shows a decreasing trend as the number of topics increases.](chart.png)