Social Search
Introduction to Information Retrieval
CS 221
Donald J. Patterson
The Anatomy of a Large-Scale Social Search Engine

Denton Horowitz
Aardvark
denton@aardvark.com

Sepandar D. Kamvar
Stanford University
skamvar@stanford.edu

ABSTRACT

We present Aardvark, a social search engine. With Aardvark, users ask a question, either in natural language or in text. Aardvark then ranks the questions to the person who may have that question, and answers it. In the past, we have used social search engines to identify a user's information, and then the right person to answer it. However, with Aardvark, this process is reversed, and we identify the right person to the right information. The Aardvark engine identifies the right questions, and answers them, by analyzing the information available, and user preferences.

1. INTRODUCTION

1.1 The Library and the Village

Traditionally, the basic paradigm for information retrieval has been the library. In a library, people find information by looking it up in the library's collection, and then finding the document that contains the information they need. However, in the village paradigm, one person asks a question, and everyone in the village responds. In a village, knowledge is shared directly — information is passed from person to person, and the more questions a village has, the more information is available.

The difference between the village paradigm and the library paradigm is that the village paradigm is much more direct. In a village, people find information by asking questions, and then finding the document that contains the information they need. In the library paradigm, people find information by looking it up in a catalog, and then finding the book that contains the information they need.

12 Aardvark

In this paper, we present Aardvark, a social search engine. Aardvark is a social search engine based on the village paradigm. We describe in detail the architecture, ranking algorithms, and user interfaces of Aardvark. We believe this to be useful to the research community for two reasons. First, the approach used in the village paradigm to develop search engines is different than in a traditional system. Second, there is a large body of research on the village paradigm, including work on distributed systems, social networks, and information retrieval.

Following the architecture description, we present a statistical analysis of usage patterns in Aardvark. We show that, as expected, in general, search engines generate a lot of traffic, and that the traffic is highly skewed, with a few queries being very popular.

Finally, in our conclusion, we look at the future of Aardvark. We believe that Aardvark represents a significant step forward in the field of social search engines, and that it will be a valuable tool for researchers and practitioners alike.
• 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
- Ranking Function
- Query Analyzer
- 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 also?

Send

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
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 \( u_j \)
- Social Connections
- Demographic similarity
- Profile similarity
- Vocabulary similarity
- Chattiness similarity
- Verbosity similarity
- Politeness similarity
- Speed match
Pulling topics out of a question

- 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
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.)
User Interface

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?

Answer

I know who to ask

Ask  Answer  Friends  More
<table>
<thead>
<tr>
<th>Question</th>
<th>Answer</th>
</tr>
</thead>
<tbody>
<tr>
<td>What fun bars downtown have outdoor seating?</td>
<td></td>
</tr>
<tr>
<td>I'm just getting into photography. Any suggestions for a digital camera</td>
<td>I've started running at least 4 days each week, but I'm starting to get some knee and ankle</td>
</tr>
<tr>
<td>that would be easy enough for me to use as a beginner, but I'll want to</td>
<td>pain. Any ideas about how to address this short of running less?</td>
</tr>
<tr>
<td>keep using for a while?</td>
<td></td>
</tr>
<tr>
<td>I'm going to Berlin for two weeks and would like to take some day</td>
<td>I need a good prank to play on my supervisor. She has a good sense of humor, but is overly</td>
</tr>
<tr>
<td>trips to places that aren't too touristy. Where should I go??</td>
<td>professional. Any ideas?</td>
</tr>
<tr>
<td>My friend's in town and wants to see live music. We both love bands</td>
<td>I just moved and have the perfect spot for a plant in my living room. It gets a lot of light</td>
</tr>
<tr>
<td>like the Counting Crows. Any recommendations for shows (of any size) to</td>
<td>from the north and south, but I know I won't be too reliable with watering. Any suggestions</td>
</tr>
<tr>
<td>check out?</td>
<td>for plants that won't die?</td>
</tr>
<tr>
<td>Is there any way to recover an unsaved Excel file that was closed</td>
<td>Should I wear brown or black shoes with a light brown suit?</td>
</tr>
<tr>
<td>manually on a Mac?</td>
<td></td>
</tr>
<tr>
<td>I'm putting together a focus group to talk about my brand new website.</td>
<td>I need to analyze a Spanish poem for class. What are some interesting Spanish poems that</td>
</tr>
<tr>
<td>Any tips on making it as effective as possible?</td>
<td>aren't too difficult to translate?</td>
</tr>
<tr>
<td>I'm making cookies but ran out of baking powder. Is there anything I</td>
<td>I always drive by men selling strawberries on Stanford Ave. How much do they charge per flat?</td>
</tr>
<tr>
<td>can substitute?</td>
<td></td>
</tr>
<tr>
<td>I have a job interview over lunch tomorrow. Is there any interview</td>
<td></td>
</tr>
<tr>
<td>restaurant etiquette that I should know?</td>
<td></td>
</tr>
<tr>
<td>I want to give my friend something that lasts as a graduation present,</td>
<td>My girlfriend's ex bought her lots of expensive presents on anniversaries. I'm pretty broke,</td>
</tr>
<tr>
<td>but someone already gave her jewelry. What else could I give her?</td>
<td>but want to show her that I care. Any ideas for things I could do that are not too cliche?</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
- Aardvark
- miscellaneous
- finance & investing
- home & cooking
- sports & recreation
- business research
Empirical Results

The diagram shows the number of questions answered within different time intervals:

- **0-3 min**: 2.5 x 10^4 questions answered
- **3-6 min**: 2 questions answered
- **6-12 min**: 1.5 questions answered
- **12-30 min**: 1 question answered
- **30 min-1 hr**: 0.5 questions answered
- **1-4 hr**: 0 questions answered
- **4+ hr**: 0 questions answered
Empirical Results
Empirical Results

- More than 2 topics: 100%
- More than 4 topics: 80%
- More than 8 topics: 60%
- More than 16 topics: 40%
- More than 32 topics: 20%