RMIUG

RMIUG's Main Menu

- Meeting Info
- Meeting Minutes
- Email Lists
- Email List FAQs
- How to Unsubscribe
- Special Events
- RMIUG Committee
- Our Sponsors
- Map & Directions

www.RMIUG.org Thursday, January 09, 2014 04:50:44 PM

May 12, 2009

Searching the Social Network: Future of Internet Search?

Select a Year

2008 Minutes
2007 Minutes
2006 Minutes
2005 Minutes
2004 Minutes
2003 Minutes
2002 Minutes
2001 Minutes
2000 Minutes
1999 Minutes
1998 Minutes
1997 Minutes
1996 Minutes
1995 Minutes
1994 Minutes

Searching the Social Network: Future of Internet Search?

Minutes of the 12 May 2009 meeting, "Searching the Social Network: Future of Internet Search?"

About 46 people attended tonight's meeting. Josh Zapin facilitated and Jeremy Kohler recorded the minutes.

MEETING SPONSORS
Microstaff (www.microstaff.com) provides refreshments, Copy Diva (www.copydiva.com) provides the audio-visual equipment, NCAR (www.ncar.ucar.edu) provides the facility, and ONEWARE (www.oneware.com)
sponsors these minutes.

ANNOUNCEMENTS

I think RMIUG should make audio recordings of the meetings and post them on
the web site. Anyone with the appropriate expertise want to do that?

Ubuntu is a free OS based on Linux--you should check out the latest version.

Boulder Digital Works is a new venture at CU, sponsored by Crispin, Porter, and Bogusky, to create a university-level set of education around doing interactive work. Ask Josh for more information.

INTRODUCTION (Josh Zapin)

The pervasiveness of Internet social communities is astounding. In particular, the growth of two of the most popular tools, Twitter, the microblogging platform, and Facebook, the social networking juggernaut, have
been the technology story over the past 12 months. According to Quantcast, Twitter's traffic has more than doubled from 6 million to 15 million visitors in the last MONTH alone. Facebook is also on a tear with more than 200 million active users, more than 100 million of which log in every day at least once. There are also reports of Facebook adding users at a pace of 1 million per day. To say that the space is growing is an understatement.

These staggering statistics demonstrate how the masses are increasingly engaging with these tools. These aren't just check-it-out-once-and-leave type of sites. These are becoming part of people's daily habit like reading the newspaper used to be or email is today. And the amount of content that's
being created around it is staggering.

What is also clear is that users increasingly use these tools to filter their Internet. Just like "classic" search engines, these tools bring you relevant information. The big difference is that you are getting it from people you know and trust. Trying to find a mechanic for your BMW? Would you
rather Google tell you which ones to go to or your friends in your neighborhood? The marriage of search within a social network is extraordinarily powerful.

Today we will explore search and social communities. We will begin to understand how social communities work and how they can be married with search to provide different or (maybe) better results.

ABOUT THE SPEAKER

Alessio Signorini (alessio.signorini@oneriot.com) is Director of Search Technology at OneRiot, a Boulder-based social search engine that finds the pulse of the web by prioritizing information based on its current popularity among the social websites. Prior to OneRiot, Alessio was a Director of Technology for Ask.com, one of the largest search engines on the Internet. Alessio is completing his PhD in Computer Science at the University of Iowa.

LINKS

OneRiot: www.oneriot.com
Twitter: www.twitter.com
Facebook: www.facebook.com

ALESSIO SIGNORINI

(Note: You can download the presentation at the following URL:
http://rmiug.org/html/minutes/2009/docs/presentation_090512.pdf)

I'm from Italy. I used to play professional soccer. Now I'm doing something a bit different.

People use search engines for everything, but now we're also using social networking sites a lot. Combined, the technology can be very interesting.

A social network is any group of people doing something together, like having a meeting. So social networking doesn't only mean Facebook. People use social networks for a lot of reasons. They evolve and increase in size. Lots of government intelligence gathering is based on analyzing social networks.

Facebook, MySpace, etc. are social network services. It's fun to connect to friends and family. But Facebook was not the first. Remember Usenet? BBS? Later came The Well, TheGlobe, Geocities. But no one really cared about those things. Facebook is much cooler.

Remember everybody had a blog? They could be searched, and people used blogs
to keep track of their lives, post photos, etc. SlashDot is a blog that became an alternative news source, and it's still huge: 900,000 blog posts published daily. A huge business. Good for advertising business. Blogspot has 5.6 million people in the U.S. visiting every day.

GROWTH OF SOCIAL NETWORKS

Friendster: Didn't work so well, but it was the first. 90 million registered, but no one going there anymore. The point was to meet similar people in your area. Friendster had declined a $30 million buyout offer from Google. Too bad--it is declining now.

MySpace: Launched by eUniverse. It did way better than Friendster. But it's now losing users. In 2006, MySpace did a $900 million deal with Google for advertising. The problem with it is that it lets you design exceptionally ugly pages with terrible colors. Just look at a random MySpace page and you'll see.

Facebook, on the other hand, is very well designed. Everybody likes Facebook. Mark Zuckerberg launched Facemash in 2003, which then became
Facebook in 2004. 100 million users are now logging in daily. It offers an open log-in platform, an internal ad network, and a public API. This allows for targeted advertising and lets other platforms tie in.

SPECIALIZED SOCIAL NETWORKS

LinkedIn is like Facebook for grownups. Epernicus is for scientists (so there aren't a lot of users). Flixter is for movie lovers, and it's very nice.

Ning lets you create your own specialized social networks. I might create one about soccer shoes. Or one for pet lovers, anything. Ning is not big. Facebook is general, with no particular topic; Ning is about something specific.

SEARCH ENGINES

Remember Altavista? It was very nice, pretty amazing for the time. It relied solely on keyword matching. But today we expect much more. We get maps, conversions, weather, flights, etc., all from the search engine.

People think search engines can answer any question. You get some very funny queries, like: "What's the mpg of the car in my garage?" as if the user expects the search engine somehow to know.

Search engines are very good at answering even difficult questions. Many companies believe in this. Google was the first, in 2001. Back then it was a search service where a Google employee would do a search for you and email you an answer later. That lasted one day only, too much traffic. They opened a proper service later and had the editors do all the work (rather than subject matter experts) but decided to close again 2006.

Today the most popular question-answering service is Yahoo Answers. It's very polished and taps the Yahoo community to write answers. Its competitors are Answer.com, AllExperts, and ChaCha. Initially ChaCha was popular: You send a question, they send you an answer--but the key was that you can do it on your cell phone. Of course now there is iPhone, which lets you find the answer yourself. ChaCha was too expensive--it couldn't make money giving answers in a couple of minutes, so then it outsourced to India and went downhill from there. Other services offer virtual visits with experts where you pay for time--like doctors, fortune tellers, etc.

Now how do you marry question-answering services to social networks? Use some artificial intelligence here. For example, if I know what your interests are, maybe I can send the question to someone whom I think can answer in a couple of minutes. Who is likely to be online right now and know the answer? Well, no one is doing this yet, so here's your chance to start a new business!

What if you don't have the right friends on Facebook to answer your question? But if I look at the searches that you do, and I can find out enough about you (the way targeted Google ads do), then I can group similar users who aren't necessarily friends and, for example, know what movies they like. Based on that, I can recommend movies you're likely to enjoy too. This way I can connect people without them knowing that they are connected.

>From your IP address we know about where you live and what kind of Internet connection you have. Google ads can track you based on a cookie that Gmail places in your computer, for example. So if you're buying a car or ordering a pizza, you'll get local results. This means that search results will vary depending on where you are when you submit the search request.

Audience question: How do I protect my privacy from sites that place cookies?

Cookies are not evil. Facebook uses a cookie to remember who you are during a session so you don't have log in every time you turn a page. I don't think you need to be nervous about cookies. But if you are, I suppose you could set your browser to not accept cookies. But then many websites won't work, especially when you have to log in somewhere--so you need to accept cookies at least for those sites.

Cookie trackers really don't care who you are--they just want to know what you do. With cookies, I could perhaps create a profile of you and use it to figure out what you mean when you search for "apache" :Is it the Indians, the helicopter, or the web server? If I tracked your previous searches and page visits with cookies I can make a pretty good guess which one and give you better results--and that's not necessarily with bad intentions.

Audience question: How do you break out of the mold that the engines have

Audience question: How do you break out of the mold that the engines have created for you?

It's a problem. But even Google will try to throw in some other stuff just to see, just in case it guessed wrong. So it doesn't always provide the same results for you.

Audience question: How do you get the same results for people in different locations, like if I'm doing a conference call?

Well, the search engine companies use the IP and you can't control that. Google would use the IP of your router, which is assigned by your ISP. There is a little application (Tor network) that people use to hide their location, but that might not work so well. On the other hand, if everyone uses a proxy server, then everyone will get the same results even if they are in different locations.

Audience question: Don't companies change the posted price of something depending on who is looking at it?

Yes, sometimes they try that, but not too often. I wouldn't worry so much about that.

Audience question: But what if someone gets hold of the information that Google has collected about me? How well is that information protected?

It's saved in internal servers, in huge log files. Google has an incentive to keep that information protected--it doesn't want its competitors to get at it.

Now let's get back to search, just knowing that there's a tradeoff: You give, and you get.

By the way, regarding those huge log files. Query log analysis is really very interesting and very powerful. It's an exciting field to get into.

RANKING URLs

Traditionally, URLs in search results are ranked by relevance and popularity (how many people are linking to your page, and who is linking?) plus around 200 other parameters and nuances are considered. This is the foundation of PageRank, which is Google's algorithm.

There are many other ways to rank pages, of course.

Now: How do you rank pages using social networks? What pages do people like me want to see? If you can match the user to the right group, then you can identify things that the group likes. We don't care who you are, rather, we fit you to a pattern that we do care about.

Ranking shared URLs

People share URLs on Facebook, Digg, Twitter, and hundreds of other sites. How can we use those? Digg, Facebook, and Twitter share millions of URLs per day. This seems like an easy problem at first, because popular URLs must be important, right? And you don't have to crawl them yourself because somebody already did that for you. So, no problem to produce real-time results, right?

Nope, actually it's very hard. Because you have to expand all those shortened URLs, and you still have to crawl them to do some verification. And there's "dechroming," where you deal with pages that have mixed content. Sidebars with related links, for example, are not part of the targeted main content and you have to separate it somehow. And that's difficult. Then you have to do context analysis, figure out how to deal with proper names in the content, remove duplicate content, etc. This is lots of work.

And you're still not done because now you have to do indexing of the cleaned up content. It's like a book index, and that's complicated. You use a tree structure for your index, rather than scanning a long list, because list scanning is too intensive for real-time application. So it's a lot of work and consequently very expensive to produce real-time search results.

For example, real-time signals like the number of diggs and tweets will just kill your database. And while you're doing all this stuff, you have make the information instantly accessible to users who are performing searches.

So many real-time engines are just filters of social network traffic. They are looking at tweets or just titles rather than the full content--because actually indexing content in real time is kind of impossible.

BETA SITES

Scoopler: Real-time search. It's filtering traffic. Cool, but not necessarily useful to intercept Tweets. It's a lot like an RSS feed.

Collecta: Same thing. Very cool looking, but not a search. The info is not scrubbed for relevancy.

There's lots of talk about real-time search. Twitter wants to do it. Google wants to do it. So does my company, OneRiot. A real-time search is delayed maybe by 45 seconds, not much more than that.

Now Google cheats to get its results faster: It up-ranks pages from CNN, craigslist, and other sites it knows are important. This allows Google to skip some of the processing steps.

OneRiot completely relies on what URLs users have shared. This could be faster than Google.

You might wonder if we can capture public opinion about something. Can we find out how many are getting the flu, for example? Well, let's look at tweets from people who are having headaches. Can you find a place where a lot of people are having headaches? That's the foundation of some of my researches. How about STDs? Ok, nobody tweets about that; but they do tweet about the drugs they are taking for those STDs they don't talk about--so we are able to figure that out too. Cool, huh? That's the research we are doing. It's important for public health services who might want to track the spread of misinformation, for example.

So here's how to be successful in search:

Find good pages
Be fast in parsing content
Study users, identify groups
Classify queries and assign to groups
Provide instantaneous feedback because delays will lose people Create an API to interact with your platform. Twitter became famous because everybody tweets from their cell phones. Twitter gets most searches through third-party connections, not twitter.com.

Q&A

Q: Twitter makes everyone a reporter. Do you keep track of how people are using Twitter?
A: Other companies are tracking this, especially URL abbreviators like bit.ly and tinyurl.

Q: Why is Facebook search so horrid? You can't find content.
A: Microsoft bought a percentage of Facebook to get access to the content.

But Facebook is not letting Microsoft access the data for search. However, Facebook is thinking about it. So is Twitter and Digg. The critical question is how do you make people pay for it?

Q: Is the quality of a search better with the combination of social networks and web search?
A: Well, when you look on Google, you get good background information. On OneRiot, however, you find out something that is going on right now. The social network gives you the "real-time" search engine.

Q: I'm raising my own rank by tweeting about myself. Does it work?
A: No, usually we can see through that. If you are very clever, you can fool

us, but most people can't.

Q: What about Google custom search engine, to search specific sites?
A: Doesn't help because Google won't crawl your site that often--certainly not as often as it crawls CNN, for example.

Q: What other sites do you use, besides Twitter?
A: Delicious, YouTube, and many others. We're adding more stuff in a few weeks.

Q: How does live searching affect Google? Does Google crawl OneRiot?
A: Google crawls Twitter. It cannot crawl our search pages. So Google will get to the stuff eventually.

Q: How do you know if what you're doing is working to get your pages ranked
high?
A: You can't really test this because it's a sort of black magic. Companies will help you with search engine optimization, but they don't necessarily know how the engines really work. For Google, you should use all the tools that Google provides, like site map, news tools, etc., and use robots.txt and nofollow to help crawlers focus on the good stuff. Your page rank will always be changing. Sometimes Google will raise it up just to see if people will click on it. If they don't, it drops down again.

Q: How about just have good and interesting content! And good links!
A: Internal links don't matter--you want to have links coming from other websites. Some people open fake blog posts just to link to their pages, but if Google finds out you're trying to game it like this, you'll get blacklisted and that can really hurt you. Sign up for Google Analytics--it's good info that will help you understand how your site is being used.

Q: Why use OneRiot instead of Google?
A: Google is perfect for finding existing information, maps, whatever. But you can use OneRiot to find out what's hot right now. What are people talking about right now in Boulder, Colorado. What going on there? An interesting event happening there tonight?

Q: How do I search my Facebook network to find out something?
A: You can't right now. There are some primitive services out there, but not too good yet. You really can't do it. We'd love to do it, to find out what's hot among people who are similar to you. But not yet.

Q: Any ads on OneRiot?
A: Nope. The model is to first make something good, then go for the advertising later.

Q: What about Second Life? Are you monitoring the chat there?

A: That 3-D virtual world was very cool at first, but they had trouble scaling up. People are abandoning it now, and things aren't looking good for Second Life. A lot of its high-level staff have left. At OneRiot we are not monitoring their chats.