

From Search to Discovery

Yezdi Lashkari

Discovery Expert Series is a series of interviews by John Furier, the founder & former CEO of PodTech Network, Inc. John has a history with Internet companies serving as Vice President of the Product Group and Strategic Development for RealNames, an Internet search and naming company, and founder and CEO of Labrador Software, a paid keyword search and navigation company. John also has experience at large technology companies, having spent nine years at Hewlett-Packard in various technical, marketing, and sales roles.

Yezdi Lashkari outlines the origins and limitations of collaborative filtering, the importance of Web 2.0, and how the commoditization of certain specific web technologies will benefit both consumers and businesses alike. He addresses the importance of blending algorithms to effectively harness collective user behavior, and the wisdom of crowds.

Yezdi Lashkari was a co-founder of Firefly Networks (acquired by Microsoft), a pioneering company in the area of collaborative filtering and personalization. Lashkari recently left Microsoft, where he played a number of senior product leadership roles, the last being a special assignment sponsored directly by CEO Steven Ballmer, focused on researching large scale network-centric computing infrastructures for thousands of hosts. This work is now driving one of the technical pillars of the post-Vista Windows release. Lashkari holds numerous patents in collaborative filtering, data protection and user profiling technologies. He received his M.S. from the MIT Media Laboratory and has three computer science degrees covering research areas ranging from artificial intelligence, databases, to collaborative filtering and personalization.

[Edited from original transcript]

We're here with Yezdi Lashkari for the Discovery Series on knowledge around computer filters, discovery, recommendations, and suggestions. Welcome. You've been around the block, you have some patents, you're a co-founder of Firefly - for the folks who don't remember those days, that was a pioneering company in the area of filtering, personalization, etc. And then they sold to Microsoft. And you were at Microsoft for a long time - since then, I think. Was it '99 you sold Firefly?

We sold it in '98 and I was at Microsoft through the beginning of 2006. So about seven years.

Seven years and you did some work for Steve Balmer (Microsoft CEO). Is that true?

I did work on - the last assignment was directly sponsored by Steve Balmer, yes, and it was researching large-scale enterprise infrastructure in this new world of network subject computing that folks like Google have pioneered, but looking at our large-scale enterprises and saying how can we apply the same techniques to our enterprise business and our server business.

Steve was very interested in transforming the Windows roadmap, post the Vista release, and gave me all the support to research it, to go influence Windows, and even make changes with his help.

Now you have a background in data; that'd be safe to say, right?

Background is in computer science, actually, just to give quick clarification. I have three computer science degrees from everything from artificial intelligence to databases to collaborative filtering personalization, which is some of the technology I co-invented at MIT with the other co-founders of Firefly.

What was that invention?

At the time, we had slightly different names for it but we actually coined some of the terminology that has now become commonplace. We actually made a distinction between at the time what we called “active collaborative filtering,” which was where you, for example, actively recommend a particular item or items to people you know will like it. And something we called automated collaborative filtering which is now sort of called collaborative filtering where a computer looked at patterns of data amongst a large community of users and their opinions or their explicit weightings or their behavioral patterns for a large catalog of items and try to compute correlations in terms of which items were more likely to be liked by a particular set of users, and eventually make recommendations to them from this large catalog.

We coined those terms because what we saw prior to that was a set of very interesting research systems in the academic literature, but they were primarily focused on what I have called active collaborative filtering and they were run on a much smaller scale.

Keep in mind this was around '93-'94 when the Internet was just happening. Tim Berners-Lee, the inventor of the Web had actually just moved to MIT a year or so ago to start the World Wide Web Consortium and we were looking at this new medium and saying you know at that scale, your circle of friends that you can actively make recommendations to and get recommendations from, and the amount of stuff that they can personally sample is quite small, but what if you could tap into the wisdom of this large community that was coming online and sampling all kinds of contact out there.

That's how the idea for collaborative filtering was sort of started. To give credit where credit is due just to make sure that the history is correct here, we all worked for Dr. Pattie Mae who is a professor at the MIT Media Lab and who was a pioneer around some of this work around sort of what we called intelligent agents and software agents, which were essentially smart systems, learning systems, or programs in your computer that actually observe what the user did. And try to automate some of these trials to get better at them. And her research group, which was Max Metral who was a co-founder and leader of Firefly along with me, and Upendra Shardanand who was also a cofounder of Firefly, and me were all working on different aspects of this collaborative filtering all in different domains. Max and Upendra working on music, me working actually on the entire World Wide Web and URLs, and bookmarks, to see if we could find interesting sites based on other people's bookmarks.

What's different now ten years later? I mean, the Web infrastructure is evolving, some people call it Web 2.0 or whatever, but now you have a lot people online, you have the same thesis, but in a new infrastructure . Talk about what you're seeing now in this area.

Well, I think a lot of significant things have changed. Just to step back for a second and contrast that with where we are today, one of the things that was interesting around the time we founded Firefly, and as you said we were quite high profile in terms of the technology we had, and some of the concepts we created, but we were, I would say frankly a little bit ahead of where the market was at the time.

You know we were in an exciting new area but people couldn't quite figure out the application of some of the technologies. I think Forrester famously did a piece on us citing us as a very innovative company and technology, but saying this collaborative filtering thing, it's never going to be big. And it really was quite - and others near the time - it was hard to predict where the web would evolve.

You flash forward to today and it's, as you say, a very different world. You know the key differences from the time when we were around is that computer power has essentially become cheap to free to utility now. You can buy cycles on Amazon as you scale out. Related to that the fact that you have these large scale architectures pioneered by folks like Google and others that expect to make crunching large volumes of data at a scale that even we couldn't imagine at the time, you know, a lot more practical, something that Aggregate Knowledge is doing very successfully. And couple that with the fact that integrating now into a website (is much easier) which is a much more style of infrastructure than actually having to deploy software on premise inside a customer's data center makes deployment and the scale that you can deal with very much more interesting and practical.

At the same time what's happened I think on the consumer side and frankly the information side is there's been an explosion in the amount of information or choices - that is buying choices or media choices, or information sources - that a consumer has the option that he can partake in.

I know you have had Barry Schwartz on prior to this or will have on the series and I think he says as well that some studies have shown that users are overwhelmed by this spate of information coming at them and the companies who succeed are the ones who successfully help users navigate through that.

In that context, collaborative filtering and technologies like it actually become a lot more interesting than just plain search of content or keyword-based search because in a lot of cases keyword-based search yields a starting point but (the) starting point is just an entry point into a navigation of information flow, and find the right items from this swath of information is something technologies like collaborative filtering can do very well. Because a lot of this is subjective and it's based on the wisdom of millions of people like you out there and collected lots of lots of data points.

So I think to sum up, there has been a huge shift in the technology infrastructure that enables us to do this at a much, much better scale than we could earlier at, frankly, much lower costs per CPU cycle than we could. At the same time there has been a huge shift in terms of consumer information and the need for such tools and such technology to help them navigate the scene. So it's almost a perfect storm and a great time for someone like Aggregate Knowledge to be in the market.

Talk about navigation issues, because this came up with Barry (Schwartz) around navigation and search. We're back almost to the same challenge we had in the early days of the web where all this new mass clutter is coming in and the user experience needs to innovate on top of that. What are some of the things that you see search and navigation evolving to, and talk about the Aggregate Knowledge approach versus others?

Okay, so I think actually that's a great question. Again I'm thinking back to our experience at Firefly and I think we pioneered a lot of great concepts, but I think we also got a lot of things wrong - we were looking at some of these things from more of an academic or computer science problem perspective as opposed to what the user may have felt more naturally to do.

I'll actually use Amazon as a great example a consumer experience that has actually used different technologies quite effectively in terms of blending search, discovery and other kinds of information at the right time in the right context.

Being more practical, all these technologies in isolation aren't really interesting to the consumer. The consumer wants something that just works in the right context, in the way they expect it to work in that context.

So let's walk through search, discovery and some of the other techniques and see where they're applicable because I don't think there's a one size fits all to be very honest. I think where search works really well is when you can specify through a set of keywords or a set of attributes say in the case of a shopping application, or potentially some other kinds of attributes what you're looking for and you have a very clear idea of what that is. Google has done a phenomenal job monetizing that model but at some point in a lot of domains, search is a starting point. And what really works once you find something that's relevant is to say show me more like this, or who else likes this and sort of navigate by opinions about that community or the wisdom of the crowd-you know, pick your favorite term.

That's where something like Aggregate Knowledge's technology, and collaborative filtering, and other forms of discovery technology really shine. So in a lot of cases, this collaborative filtering fails in certain deployments, and even in our case we learned this the hard way where we put it up there and said this is a recommender system. And recommendations without context are frankly not that useful to consumers. If I just say recommend stuff for me, that doesn't really make sense - it has to be customized in the context of what they're trying to do or what they're looking at. And I think this is where I go back to someone like Amazon who still is the gold standard of how to get these technologies into a very compelling experience for what they do.

Search can be looked at as a starting point but in a way from a user perspective and it could be the failure point because if I can get information and navigate a solution for me, without touching search, that would be the ultimate. I mean you don't have to search for something; I'd like to have it presented to me kind of in advance.

That's a very interesting – it's a proactive thesis and I'm not sure I would quite say it the way you did, but let me sort-of define it a little bit. I think search has an important role, especially when it embarks on retrieving something very specific from the sea of information out there. I think that's the role Google plays really well. But if you looked at where Google hasn't really succeeded, and you look at their shopping application for example, or even YouTube, where it's a lot more subjective in terms of what's interesting, and it can't be quantified absolutely. The success of YouTube stems from the fact that users have their own stake in the creation of the content or the ownership of the site. But it also is something that catches your eye and being able to surf just by “show me more like this.”

I think in a lot of domains where what the consumer is looking for isn't easy to specify very exactly, whether it's by keywords or other forms of search terms, but it's more like “I know it when I see it... I can't explain what it is, but I like this.” If I could just point to that token and say “show me more like this or who likes this, what else do they like,” that is a very powerful metaphor for navigation and I think the two are complimentary so I don't actually believe that it's one or the other. I do believe they're complimentary and they have to be – again go back to Amazon as a great example of how this can apply.

They have to be done at the right time and the right context for both of them to be really effective. I don't think you can do either one of them completely, and I think there have been fairly rigid approaches from both sites, maybe there's this Google search algorithm that will solve everything or this uber sort of discovery technology and I don't believe that's true. I think the consumer really doesn't care what's behind the covers, there is a natural way they want to express their need for finding something and discovering something and the companies that are going to be successful are the ones that can surface that in the right context to the right consumer experience. I think what the consumer naturally expects is super critical.

You mentioned Aggregate Knowledge; there's other different approaches out there, there are many different classes of algorithms like personalization, item-based filters, Bayesian inference – what are your thoughts on the Aggregate Knowledge approach of high-volume testing of each algorithm, blending the results for the best ROI and compare that to other approaches.

This is a great question. Let me give you a little bit of my background. What the Firefly team was chartered to do was build what became Microsoft passport, the online identity system, and now is Windows Live ID. But I was actually asked to go to the server platform, the e-commerce platform, and build their personalization and user profiling technology platform for e-commerce, which became Commerce Server 2000 and its follow-on products.

One of the things that I did there was actually work with the Microsoft Research Group looking at the Bayesian technology they had, looking at the collaborative filtering technologies we'd invented and saying, "Which one should we put into the box?" And the conclusion we came to is in a lot of cases it's a bit marginal which one works better in terms – and there are certain applications where some technologies are more naturally suited because of the data that they collect and the way they utilize the data where as others – for example, collaborative filtering is much more suited to very efficient navigation of a large catalog while Bayesian might be much more suited to weight-targeted marketing based on pre-computed classes or pre-computed feature sets.

But really at the end of the day the consumer – I do tie it back to the consumer experience and you have to test that past different ones and different dimensions and see which one seems to work.

We actually tested a number of collaborative filtering algorithms at Firefly. In fact we ran a research team of a bunch of very smart MIT folks as a subset of my responsibilities at Firefly just to prove the algorithms. One of the challenges we always had in a lot of these cases was how do you measure effectiveness of these things? How do you measure whether your algorithm is really the best?

In some sense a lot of recommendations are subjective, so how do you measure the quality of the recommendation? You can do basically some notion of classical machine-learning where you take out a sub-set of answers, try to predict them and see how close you came, but this is a moving target. So to go back to your question, I don't believe that one technology, one algorithm, works in all domains and all applications. You have to try them, but you have to try the one that's best suited for the kind of data you can collect in the application and the kind of output you want to spit out for them in terms of do you want to do a quick computation based on a set of correlations like the nearest neighbor, or do you want to do essentially some kind of a clustering which is then used to create prototypes or stereotypes for creating essentially certain segments of users that you want to then apply certain kinds of marketing promotions to.

I think they actually depend on the application a lot, And again going back, I just want to make sure that this is sort of along the lines of what you were asking.

There are different approaches, some recommendations don't always fit, and it's the classic contextual behavioral challenge that people have. When I was interviewing Barry Schwartz, he talked about an example on the brick and mortar side where a chain in Seattle presents information as a surprise for the shopper and their customers have a high satisfaction rate because they go in to run an errand they end up getting a surprise. That's different approach than just having those recommendations. Aggregate Knowledge takes a little bit different approach which piques my interest around the algorithms that they're using and what's the best for a site to deploy.

In some sense if you think about what they're doing, I think it's almost the right approach, I mean even if you look at compared to someone like Google who's constantly tweaking their algorithms is the web and consumer trends and application domains are conveniently shifting and now you've got just such a large body of data out there you can actually live on real time live experiments and see what's working and tweak it on the fly. Which is something that both from the computing infrastructure perspective and just the applications perspective, we really didn't have the opportunity to do at Firefly ten years ago. And so I actually think what these guys are doing is really smart which is you constantly have to keep your finger on what seems to be working and what not.

I have had tools to figure out whether you can do better so I know they do a lot of A/B testing for example and it's super critical. You can't rest on a particular algorithm as being the ultimate one because the landscape is shifting the way consumer navigating media is shifting on a continuous basis.

Today it's easy to do - you have JavaScript. What do you think about the current generation of a few lines of JavaScript to integrate and what that means for the customers of these systems? Talk about the ease of deployment. Deployment was a big issue when first generation recommendation engines were out there.

One of the reasons I got really excited and agreed to help advise Aggregate Knowledge was I think Paul (CEO of Aggregate Knowledge) and I shared a lot of the same vision of where we thought this technology could go in this current landscape but also I liked their story. It's a very low-risk story for a customer to say "Look, I'll try this out by integrating a few lines of code my HTML and see how it's doing." You couple that with a pay-for-performance model such as the one that they've been pushing - what's the downside, right? If you don't like it, yanking it out is easy. You don't have to take a large up-front cost to deploy, test, and operate this thing in your data center and essentially it's continuously being updated because the Aggregate Knowledge guys are incented to make it better. That's how they get paid. It's a win/win situation all around for customers, actually. And I wish we had, you know, some of these technologies and this infrastructure in place 10 years ago because it would have made deploying so much easier.

So the trend is lightweight integration via JavaScript to embed and pushing the heavy lifting on the cloud side, which Aggregate Knowledge does?

I think increasingly utility based models for certain key functionality - whether it's actual compute cycles or recommendation hosting or even we've seen this with productivity like e-mail and document hosting - is becoming something that is becoming widely accepted both by consumers and frankly by enterprises. So I do think the trend is the right one.

That being said, there will always be certain customers that, for reasons of either proprietary information or control or certain other reasons, want to run the whole thing in-house, and those will be typically much larger customers who've got in-house programmers, and IT staff, potentially Ph.D.s who might even develop a lot of this recommendation system off the shelf and customize in a very specific way. But I think for the large set of customers out there and growing, what Aggregate Knowledge and folks like them provide is a very lightweight and very efficient way of giving the infrastructure of the Internet now and broadband to actually trial this or the deploy the functionality, see how it's working and decide whether this is a core part of their business very quickly.

Just a final topic area, talk about the state of the software business. Collaboration is a big part of the social scene right now in terms of social media, social computing, and more people are connected, people are sharing, this is a phenomenon. Share your thoughts on software environment and the role AI plays.

Yeah, it's a very broad question. I don't know if I can do justice to it. Well, I think you have a lot of different people, and people frankly who have had much more illustrious careers than mine. Folks like Michael Dertouzos who was the head of the department of Computer Science at MIT (who) passed away a few years ago, but I think the large scale vision and what you're seeing is the infrastructure of information which the Internet and the web has sort of essentially sparked and now become a global standard, is going to be just expected to be available through any device anywhere you go. Interestingly enough this new generation of consumers, you have what I call "Generation Me," sort of the 18 to 24-year-olds and the ones coming after them, have only grown up with the Internet and with cell phones and been continuously connected in social networks.

They've never been in a world where they don't have connectivity, as well as social connections all the time so they are living with the information actually very comfortable in the sea of information. I think what you're going to see is basically an expectation that wherever I go to with any device I will be able to not only access any information I want but actually be able to access the information I need in that context to make the right decision and be able to essentially do a lot more of my living in a virtual way. You know, in terms of obviously a lot of my life is organized today using my digital calendar which published the web and email and instant messaging, but increasingly more so, I think, you'll find you know folks such as retail environments becoming a lot more blended with the Internet with mobile phones and point of sale technology. You'll find increasingly—we've already seen this trend with things like Second Life, people sort of the blurring the line between online and the real world being so fluid that, a lot of cases, what is real and what is virtual will almost be a - not the right way to put it - it's probably just a detail in some ways.

I don't know what that means for artificial intelligence in the classic sense, but I think we're at the stage here where you can say computers are pretty damn smart in terms of looking at large patterns of data and being able to pretty uncannily predict or at least make some good guesses about what might be relevant to me in a given context.

I think technologies like recommender systems are a small part of that, but just the fact that the computing infrastructure allows me to access to any kind of data from anywhere, have them synthesize that with a large dataset very quickly just means that computers are a hell of a lot smarter and a hell of a lot more useful for my everyday life than they were even 10 years ago.

It's something we take for granted. It's kind of scary in a way because I remember in '95 when the web was still new, we had to show venture capitalists what this Mosaic thing was, and now you know you don't think twice about turning to the web for everything.

And this is not just information, this is about transactions, this is about your love life, this is about connecting with friends all over the world. I think increasingly this virtual infrastructure is just going to be all around us, it's going to be like air and you'll access it through all kinds of interfaces and devices including voice, including gesture and touch which are things that are still a little bit off in the horizon but a lot of great research going on in large universities and corporate environments.

And it will be something that for the generations that come five years from now will just be the way of nature? It will be part of the environment. It won't be something where you have to take a laptop out, you can just talk to the building. And the building will understand you, decode your request and identify you and then give you the information or do the appropriate action and you won't think twice about it.

Final question for you - wisdom of crowds, collaborative filtering, collective user behavior, collection of data, and it creates better choices for people. Where do you see the next couple of years in this area emerging?

That's an interesting question. I think what folks like Aggregate Knowledge and other people like them in the market today, what I call second or third generation collaborative filtering systems, are doing is they are creating large sets of databases of behavioral data across various domains, and I think what becomes really interesting is when you can take that data in a very natural way while preserving privacy for the consumer, because that's of paramount importance in something like this. You know giving consumer control over their privacy or their profile but augment that simple, natural task that they do and make them just more easy, whatever they are, today folks like Aggregate Knowledge are focusing primarily on the online environment because the infrastructure and the ease of deployment are - and the amount of data you can collect has tremendous value.

But you know we still live most of our lives in the physical world, right? And when you can take that and couple it with what you can do in the physical environment or in environments other than e-commerce or sort of content related which is again where some of the systems are focused on today, but just leave it into a part of my everyday interaction for example into say, Facebook, basically being able to look up some of my friends and figure out who I should invite for this particular kind of theme party.

Now some of this I know because they're my friends, but what about friends of friends? Being able to leave that and to just very simple interaction the transactions that I do, but in a natural way, I think that is the future of some of these technologies. The challenge again is going to be how do you make it natural for the consumer so that it doesn't feel like a hokey step or something that is outside the realm of consumer behavior.

I think when you can do that, and you can use the existing infrastructure that's when you truly also start blending online to offline in a way that the consumer doesn't think is artificial but seems completely a logical extension of what they've been doing all along.

Really the intersection of online and offline behavior and data and value is going to be a key thing. That's exciting. Well, we'll definitely keep in touch and you've had a great career so far and you've done great pioneering work in the area of filtering and computer science and obviously you are advising Aggregate Knowledge. We really appreciate your time chatting.