PhD Comprehensive Exam

Terrell Russell
University of North Carolina at Chapel Hill
School of Information and Library Science
Friday, January 29, 2010

Question 5 – Marchionini

You make the case for contextual authority tagging as a mechanism for inferring interests and expertise and you discuss methods for operationalizing CAT. You also argue that as people use more social media to project themselves publicly that their personas will coalesce, which presumably will support better expertise finding. 1. Briefly describe why and how this persona convergence will take place. 2. Discuss the underlying assumptions (e.g., most people will give honest tags and not manipulate the system) necessary for expertise finding based on public CAT to eventuate. 3. Select two of the underlying assumptions and describe what might be done to work around them to support expertise finding in they do not hold.

Response

1 Introduction

I have proposed a line of research that gets at the question of whether a group can know what an individual knows about. I chose to work at the level of the professional organization as it provides the best environment for considering issues of expertise and skills. The underlying assumption of the preliminary research is that if it does prove interesting and useful, then the same methods could work at a much larger, public, or even global scale. This essay will look at this globally public version of Contextual Authority Tagging (CAT).

In section 2, I talk about why and how the convergence of our public personas will take place. In section 3, I name and discuss a few assumptions about what needs to be in place for a public CAT to take hold and function successfully. In section 4, I pick two of these assumptions and think through what we would need to do if they were not present.

2 Convergence

As we continue to move into a fully networked and fully read/write future, the democratizing effects of everyone having a voice will continue to bear fruit. On the positive side, we have a growing public discussion around most things that matter (politics, religion, science). On the other hand, we are beginning to drown in an ever-expanding flood of opinion and rhetoric. One of the additional effects of all this bidirectional media is that many people are leaving public traces of their interests, activities, and opinions in places and in volumes that were never possible before. As a by-product, vigilant actors (computer programs) can begin to keep track of what an individual says and does and the things they like to discuss. Perpetuating multiple distinct profiles will become harder and harder, if not impossible, in the near future.

2.1 Why

First of all, I think that living in public is something we have not fully considered as we move forward with our increasing technologies and propensity for living in cities rather than in the country. We are moving ever-so-quickly into danah boyd's *Networked Publics* and will have to deal with replicability of the data we generate, the scalability of our reach and its hidden audiences, the persistence of our networked activity over time, and the searchability and recontextability of all that is affiliated with our accounts. The sheer amount of data and information will quickly become (has already become) too complex to successfully keep separate and we will not be able to keep track of who recorded what and how that could potentially be aligned with other snippets of information we left somewhere else. Our projected personas will collapse upon themselves when interrogated with any vigor. Keeping separate identities will largely be moot as our algorithms for pattern matching and similarity detection continue to improve.

Corporate databases are already keeping track of much of this type of information - but focused around purchasing and demographic data. This will continue to happen out of the line of sight of the public. As more published, web-born data is created and lives online, I guarantee it will gladly be ingested and used to improve the already fairly comprehensive datasets these large companies are keeping. Additionally, as the information is being generated *on* others' sites, out of the control of the individuals doing the creating, the individuals have no real control over the future broadcast, sale, or integrity of that data. By moving data collection and publication into the public, people may prefer it as going public affords them a greater sense of control over what is being seen. Knowing what information is out there about oneself is better, even if knowing who is seeing it is still a mystery.

One last reason this identity collapse will occur is that people like to talk about one another and they like to keep score. Pageviews, Facebook friends, and Twitter followers are an indicator of our need to report status and importance. I suspect those people who are trying to keep separate personas will be unsuccessful for no reason other than good

visibility into lists of friends and activities will uncover hidden patterns that give away some secrets. This could be construed as a chilling effect, but I doubt the spirit of those who are trying to keep things separate will really be suppressed – rather, they will embrace who they are and move forward publicly once the option of keeping things separate is off the table. If a separation was being attempted for subversive or illegal activity, we could argue that reducing that possibility is a good thing as well. Illegal activity will go further underground and anonymous.

2.2 How

As to how this convergence will happen - I think there are four main points.

First, it will not happen all at once. There will be first movers, early adopters, a mainstream, and laggards. Like most things technological, the nerds will move first and have inordinate amounts of influence on the design and functionality of any products involved in the convergence of identity. I think this is already happening and can be seen in the desire for single-sign-on across technology sites as these individuals began to have an explosion of accounts across the web. The first attempts at this technology were proprietary (MS Passport), then moved more open (OpenID) but was ugly/technical, then grew some proprietary polish (Facebook Connect, Twitter Connect), and will hopefully swing back to being open and easy to use. As this process continues to evolve, more of the mainstream and laggards begin to play with the same tools as the early adopters.

Secondly, there will be great concern about the loss of privacy and lack of control in how we are able to segment our activities and interests on the web (and therefore, the world). This is a natural response and one that must play out every time a new technology begins to take hold. I suspect we will see requisite business models appear that are built on the premise of limiting the amount of information being collected and passed around about its customers. This too has already begun to happen as we have companies like Reputation Defender and claimID positioned to help the individual have more of a voice in the conversation happening

around our new converged identities.

Third, I think the convergence of identity will be fought by those who do not wish to see it happen. Data will be hacked and forged and changed and gamed by many parties who wish a system of identification were not taking root. As identity (and expertise) data become more visible and widespread, things will break a few times. But we will learn how the systems failed, and they will become more robust. Going out on a limb, old people (change is hard) and people conducting illegal activities (bad for business) will be most against this collapse of personas.

Lastly, it will finally hit a tipping point and we will realize that normal has shifted. It will seem strange if you cannot look up a potential hire's previous work history and evaluations. It will seem strange if you cannot search the web for a quick background on the reputation of a restaurant (oh, wait...). It will seem strange if we cannot immediately have access to a Whuffie-like score that represents the good deeds or the goodwill a person has among their friends and acquaintances. When normal moves, it will have moved because there was value in the type of information being provided. Going backwards will not really be a possibility and, at that point, I suspect that the convergence of public personas will be fairly complete. 15 years? Hm.

3 Assumptions

There are some basic assumptions that I have about how the world works with regards to CAT and its potential success. I think each of the following assumptions are necessary for public expertise tagging to take hold.

3.1 Value

Largely, we live in a market economy. Things that have value are worth time and money and those things that are not valuable become commodity and go down in price. I suspect that

knowing what people know is of great interest, importance, and value to many firms. I think that the reason we do not already see such a marketplace is that the information on which to determine this value is largely hidden (because it is extremely expensive to uncover or it is taken into consideration by other costs). If the cost associated with *seeing* what someone knows about is reduced, I expect there to be a shift in power to those who do have expertise that people can see.

3.2 People are Social

Humans are social creatures and are interested in one another. Given the fact that we have language, we convey this interest and curiosity through gossip. We are the most interesting creatures to us! Given a set of tools to foster this gossip, I suspect many will begin to use them regardless of a lack of full understanding of the possible ramifications or the possibility for abuse. I assume that because the object of interest is other people, a tool like CAT would be interesting at face-value. If it happens to be useful, then that just makes for a stronger case.

3.3 Open Code

A system like CAT would need to be open and visible. The code would need to be available for inspection before it would be trusted. As a distributed system, each node would be running its own version of the code and would require built-in cryptographic signatures and other failsafes to provide for the integrity of the infrastructure as a whole. That said, the open code would be speaking via an open protocol. The passing around of tagging data would need to happen in a standardized way that is well understood and well-studied. No one entity can be in control or the integrity of the entire system would be in question.

3.4 Open Data

On the heels of the open code and open protocols, I think that the data generated by a system used to determine reputation and trust must itself be open and capable of inspection. A robust system of checks and balances can only happen when the data itself is visible and verifiable by third parties. If the data is visible and verifiable, then others can vouch for its "goodness" and can be safely replicated and rebroadcast. Open data also allows for clearer chains of provenance and custody. Clear lines of history are critical for understanding how data came to be what it says it is.

3.5 Good Data

In addition to open data that is verifiable and consistent, we must assume that the data itself is meaningful and good. This is really an assumption about the people rather than the code, though. Users of this system would need to be largely well-behaved and using the system as intended, or at least in ways that are not malicious and designed to undermine the usefulness of the system for others. All systems will have some free riders and bad actors among the ranks, but in large part, they can work as long as some threshold of the participants are following the rules.

3.6 Legal Protections

Another assumption is related to our system of existing laws. An automated system will always break in certain ways, and good systems have failsafes that catch exceptions, or error modes, and deal with them gracefully. A system like CAT would need to allow for the law to take over when issues surrounding libelous statements are made and propagated or when falsely impersonates someone else in order to manipulate the system's output. Issues that cross international lines would continue to be complicated and complex.

3.7 Tipping Point

As any system gets off the ground, there is a point in its development and deployment whereby enough people are interested to keep its future secure and usefulness above water. I assume that there is a point for CAT as well, and that before that point is reached, very little interest will come its way. Additionally, I think there is a natural distribution among individuals that would be using CAT and that like most other social systems, there will be celebrities with a vast amount of information about them in the system as well as people about whom there is very little information published. Those who have very little information published about them will see little value come back to them with regards to being sought after for their knowledge and expertise. I also foresee people who are more social and extroverted having an advantage in a system like this as they would naturally be interacting with more people and therefore having a greater chance of being evaluated in a global reputation system. That said, I think that once a bootstrapping horizon is crossed, a system like this would remain useful for many people for a very long time.

3.8 Analytics and Filters

This is one of the most interesting assumptions. I assume that for CAT to be useful, in addition to having met a threshold of activity and data, it must generate a spectrum of tools for analysis and filtering in order to be useful. After a while, there will be so much information in a system like this that ignoring most of it will be normal. Filtering could be done based on time or tagger or topic. Filtering could even be done based on groups of people or organizations or by geography. Analytics could provide insight into decay rates or areas of recent skill acquisition. I think these are essential tools that need to be built or the flood of tagging data will be largely unapproachable and indecipherable.

3.9 Anonymity is Noise

One last assumption is related to the anonymity afforded by a system of this type. I think that anonymity is an important aspect of any reputation system. It is a necessary potential avenue for feedback and has its role when the balance of power is unequal and to speak with identity could be dangerous to the speaker. But, I also think that because of this, most anonymous speech is useless and noisy and should be ignored. Our newspapers and magazine sites that still employ anonymous comments are impossible to read and provide little to no useful commentary on the story at hand. As I said just prior, filters would need to take this into consideration and allow for users of a system like this to largely ignore anonymous tagging as I assume it would be mostly junk. Anonymity should be a possible avenue for a tagger, but it should definitely not be the default, as it is unattributable and therefore potentially undermines the credibility and trustworthiness of non-anonymous speech.

4 Work Arounds

In the previous section, I listed nine assumptions that I believe are necessary for CAT to function well in a distributed, public environment. Some of them I feel are a given and without much room for disagreement. Others, I could be wrong about. I'll pick two here.

4.1 Bad Behavior

If CAT is a reality and is rolled out and has many participants and is found to be generally useful, it will begin to draw the attention of spammers and other misbehaving actors. If it is truly interesting (read: profitable) to spammers, then the majority of traffic within CAT could quickly become spam. This is what has happened to nearly all channels of information that hold value for users in that they give it their attention. If a plurality or majority of traffic in a channel is found to be spammy or malicious, then the attributes of having all information signed becomes of paramount importance. Part of why email spam

has become such a problem is that the messages are nearly all unauthenticated, even when they are legitimate (that and is inherently a push technology with virtually zero cost for pushing). This makes it extremely hard to identify real email from fake email. If email is mathematically signed, it is very easy to determine which messages came from known entities and which do not. As CAT would, by definition, have tagger information on all tags, this should largely be an issue of allowing spammy information to exist, but then filtering it accordingly on the display side. Having (identity) authentication built into the infrastructure of CAT, bad behavior can be both tolerated and ignored. A happy side-effect is that if bad behavior is not disrupted, it usually stops.

4.2 Proprietary Code

I stated earlier that open code would be an assumption for a healthy ecosystem to grow around CAT and trusted tagging in general. If the code is not open itself, then the path to trust and confidence in the system may be longer, but I still think it is possible. The other key pieces of this puzzle are that the protocol and the data are still open. If those two are still the case, then even if the code that generates the open data and sends it around via an open protocol is not visible and available for inspection, other actors can still evaluate its performance and network behavior. In fact, I think that proprietary code may actually be a good thing, in that it may push forward the level of innovation. But in the long term, open code will be the thing that creates the trust and confidence in a system that is trying to generate reputation. Visibility is key and without it, any confidence and trust is only one misdeed away from being torn down and destroyed.

Also, with the same arguments applying in this case as with the spammy behavior mentioned above, if a proprietary actor begins to misbehave, the rest of the network can easily ignore his activities and his data. Spamhaus is one example of how the network has developed a trusted clearinghouse of bad actors on the network. If bad actors become an issue, the good actors will band together in affiliations and federations to better understand how

to trust one another. And again, that trust will happen much more quickly and with greater confidence, if code is openly shared and debugged.