PhD Comprehensive Exam

Terrell Russell
University of North Carolina at Chapel Hill
School of Information and Library Science
Tuesday, January 26, 2010

Question 2 – Wildemuth

How far can the Delphi technique be stretched?

As noted in your literature review, the Delphi method was developed as a forecasting technique in the late 1950's and early 1960's. Since then, it has been used in a variety of research studies, but it has also been criticized for its methodological weaknesses.

> What are some of the criticisms that are most likely to apply to the way that you plan to employ the Delphi technique in your dissertation research? As you develop your plans for your dissertation, how will you address these criticisms?
>
> Looking at this methodological approach in a more positive light, what are some of the attributes of the Delphi technique that make it most promising for your dissertation research? What is your rationale for adapting this technique for use in your dissertation research?

Based on your preliminary plans for your dissertation, you will not be using the Delphi technique as a research method, but instead using it as the "kernel" of a system supporting Contextual Authority Tagging.

> What methods will you use to evaluate whether this approach to gathering and aggregating participants' tags of each others' expertise is effective or not? What criteria will you apply in this evaluation? How might you operationalize those criteria?

Response

1 Introduction

The Delphi method was created during the Cold War to assist the US Military in determining the nuclear capabilities of the Soviet arsenal. Determining such a thing required vast amounts of information that was actively being hidden from the US Military. Delphi allowed an anonymized, iterated, aggregated opinion to be collected and analyzed from those who *did* have some insight.

Of course, since any answers that come from a Delphi study are inherently fuzzy and without attribution (only the researcher knows who said what), these answers must be considered carefully. However, the point of the Delphi is to get at information that is otherwise hidden or distributed too widely to otherwise be seen. When this is the case, applying a Delphi could generate the best insight available. Measuring opinion is inherently messy, as there is no objective truth and no yardstick by which to measure success. By looking at consensus, or less demonstrably, a plurality of opinion, we can gain some understanding of where a group stands on a question.

Applying the "kernel" of the Delphi to expertise location comes with the same caveats. In section 2, I address some expected criticisms that could be leveled at my proposed dissertation research. In section 3, I discuss why I think that the Delphi is the best foundation upon which to construct Contextual Authority Tagging. And then lastly, in section 4, I talk about how I may evaluate the effectiveness of this proposed technique.

2 Expected Criticisms

Contextual Authority Tagging aims to generate sets of words describing individuals' areas of expertise. The main justification for this type of information is that current systems are generating expertise descriptions from only the subject's point of view. Automated

algorithms are currently looking at documents produced, email, and attention activity – but all from the person being evaluated. I think there is additional value in hearing what others have to say. But of course, getting people to talk about people is inherently less clean than an algorithm running on a stack of documents.

The standard criticisms that have been leveled at the Delphi include two major types, concerns about the participants, and concerns about the data that is generated. Regarding the participants, established weaknesses consist of concern about the level of anonymity involved, the methods/criteria surrounding the selection of experts, and the demographics of the selected participants. Concerns about the generated results revolve around the lack of statistical tests to show significance and/or rigor and therefore the lack of predictive power associated with any/all of the statements that come from the participants.

With regards to the proposed Contextual Authority Tagging, some of these are more salient than others.

2.1 Anonymity

Any concern over the **anonymity** of the participants or the attribution of the tags is reduced to the security issues around the database where the information will be stored. For research purposes, I plan to store both the "tagger" and the "taggee", but this would not be strictly necessary if plausible deniability was of due import. Further concerns over who said what are relegated to the realm of the social – the scope of which is beyond the aim of this research. I am assuming that by making these expertise tags visible and available for discussion, some stories regarding the provenance and justification of the tags will be told. Truly secret information should remain secret, regardless of the availability of a tool or exercise like CAT – but that is an issue between those who have secrets and those who know the secrets.

2.2 Selection

I am also expecting to hear feedback regarding the **selection** of participants of the form "friends/colleagues are not experts." I posit that they are expert, within the context in which the experiment is run. Some information about a participants' areas of expertise will surely be beyond the purview of the other participants involved, regardless of the environment in which the experiment is run. That said, I am limiting the research to be run in professional or collegial environments where intellectual activity is the main type of interaction between participants. I am explicitly avoiding, at this time, groups that could be construed as family, social, hobbyist, or athletic. By sticking to offices and workplaces, I expect that the types of information generated by CAT to remain largely "on topic" as that's the nature of the majority of interactions between the participants. Additional "off topic" information would be generated and displayed as well, but it is expected that these would be limited in scope and not extend much beyond what is commonly discussed at work already; the participants will continue working together after the experiment is complete.

2.3 Misinformation

There will probably be concern over the possibility of **negative information** or **false claims**. These two concerns are important and deserve attention. I suspect that non-normative behavior and aberrative tags will draw attention quickly. This is no different from unprofessional language being uttered or a physical disruption in the workplace – it is quickly noticed and addressed. Negative tags will largely be disincentivized by the positive phrasing of the question being asked, "What do you think this person knows about?", and "What are this person's areas of expertise?". Answers like "being a jerk" would stand out and not be corroborated by others over time. That said, if it *was* corroborated and voted up by others, this tag arguably is doing a service to the community by making it apparent to this participant that they are not viewed as helpful by a contingent of their peers and co-workers. This could, arguably, lead to better behavior on the part of the tagged.

2.4 Coverage

Another concern that could be leveled regards coverage of the generated tags, and the fact that there would remain **hidden information** not captured by this technique. I agree, but do not see that as a limiting factor. I assume that humans will always hold some information to themselves – and I encourage that. I also think that the anonymity provided by CAT will allow more information than is currently being put on display to be captured and propagated around. I think having total information would be a horrible thing. I also think that having a place for anonymous speech is important and that it sometimes brings potentially fascinating and useful information to the fore.

2.5 Statistical Rigor

Regarding the **lack of statistical measurements** and tests to determine the significance of the findings rendered by classical Delphi, I feel CAT can be claimed as immune. The nature of Delphi is that it results in a set of findings or opinions that have been deemed "convergent." The weakness of these findings can be attacked from a predictive standpoint, but as I intend for CAT to be run continuously (if implemented beyond my dissertation research), the notion that a test was not conclusive or that there is no test is a non-issue as there are never any final "findings." The co-workers will take what they want from the information and use it accordingly. I see CAT being a piece of reporting/learning infrastructure that allows other tools to be built and used around it for decision making. Making the opinions of people visible should create more opportunities for discussion and reduce the chance for misunderstandings.

2.6 Loss of Control

I also expect to see some pushback from (potential) participants regarding their **not having a say** in what is being said about them and the fact that this information is being published

for others to see. My counterpoint is that this is already happening, everyday, all around us. People gossip. CAT will just bring this information together, aggregate it, and show it publicly. Those who are good at what they do, and know their stuff, will be rewarded. Those who have not convinced their colleagues of their areas of expertise will have sparse data to show for it. This is not as much a privacy concern as it is an issue of control. CAT, I agree, definitely moves the control of defining ones areas of expertise away from the individual and towards the group. But I also think that this is a good thing and something we need as we begin to live in an ever-connected, online environment where notions of identity are not as ingrained and well-understood.

3 Rationale

In this section, I hope to explain the reasons why the Delphi is the method on which I have chosen to model my research. First and foremost, the Delphi consists of three main criteria: expert opinion, anonymity, and iteration. Secondly, it was designed to find a consensus among the expert opinion.

Contextual Authority Tagging is designed to find the aggregate opinion of a peer group with regards to each members' areas of expertise. It is designed to be a continuous process (but will be modeled in my dissertation as a stepped process) and provide insight into the weight and diversity of knowledge of a group's members.

I have mapped the salient attributes of the Delphi method to CAT in Table 1. The attributes map in a straightforward manner and should provide some clarity around the roles to be played in my modification of the Delphi. I will speak more about this table in section 4.

As the Delphi was designed to evaluate subjective opinions, I have chosen it to help me formulate a methodology to look at subjective opinions specifically around expertise. The irony of looking at expertise with a method originally designed to use experts is not lost

Attribute	Delphi	CAT
experts	selected experts	professional colleagues
opinion	answers to specific battery of questions	areas of expertise
anonymity	protected by researcher	recorded in database only or not recorded
iteration	3-5 rounds	4 rounds (continuously)
independent aggregator / synthesis	researcher(s)	software/database (simple counting)
consensus determined	researcher(s)	third party blind raters (Mechanical Turk?)

Table 1: Mapping of attributes: Delphi and CAT

(or intentional), but I do think the thinking holds up. An individual's colleagues spend more time thinking about what that individual knows more than probably anybody else, outside of the individual herself. They are uniquely situated to evaluate the question around the individual's areas of expertise – and therefore, I'm considering them the equivalent of the selected Delphi experts. Traditionally, this type of expertise evaluation has been done either solely by the individual (via her résumé) or her boss (in a letter of recommendation or reference). I hope to add a potentially useful voice to this duo.

4 Evaluation

Evaluation of a set of subjective opinions is no small feat. As there is no objective truth by which to measure the effectiveness of this technique, I am proposing a similarity measure.

Contextual Authority Tagging will be composed of a series of rounds of evaluation via custom software running in a standard browser. Within each round, a participant will tag their own areas of expertise, and then the areas of expertise of each of their group members (colleagues). A group comprised of 8 participants will require each participant to enter tags a total of 9 times each round: once to tag themselves, seven times (once for each of the other participants), and once again to tag what they think the group will say about them. I have been referencing these as A_A, A_B, and A_A^*, respectively.

At the end of each round, for each person (A), the system will have a cumulative set of self tags (A_A), aggregated group tags (A^*), and meta tags (A^*_A). To generate the similarity scores, when the data collection is complete, I will ask a multiple of third party raters to look at these three measures, pairwise across different rounds for each participant, and rate their similarity on a seven-point Likert scale ranging from Not Similar to Virtually Identical. These third party raters could be recruited and trained by myself or recruited and tasked through Amazon Mechanical Turk.

As time passes, the similarity of how an individual tags their own areas of expertise and how the group tags that individual's areas of expertise may become more similar (a simple ANOVA should be enough). If this is the case, alternatively, I would expect to show that the same thing cannot be said for sets of tags that are describing different people (the null hypothesis).

As the gold standard for what a person knows is generally taken today as the self-authored résumé (with additional input from former bosses when necessary), this similarity measure will determine whether the group's opinion holds up. If it does prove similar, we will also gain some insight into the speed at which the convergence happens (both in time, and in the number of taggings). Alternatively, there could be a persistent gap between what the individual says and what the group says. Perhaps an individual is never "validated" by his or her peers (or read the opposite way, the individual is ever-delusional when it comes to their own knowledge/expertise).

I also plan to ask the participants about their levels of satisfaction in using this technique to evaluate their own areas of expertise, as well as that of their colleagues. If the participants report that they learned things about one another that were insightful and interesting, then I would count this experiment as useful, but not necessarily a success. If they report that the information gathered by this technique was useful and trustworthy, then the experiment is a success. This bar of success would potentially qualify CAT as a viable business opportunity as I suspect it could be shown to be cost-effective as compared to existing products/services

that generate the same type of knowledge/skills data.

Another indicator may be whether the group tended to use a set of tags before the individual or whether the individual self-tagged in a certain way and then the group followed suit. Determining patterns around this type of behavior may prove predictive in some manner down the line. Leading indicators, while relative, could still provide insight into future activity.

While these evaluation methods are all based on perception, I feel that when measuring opinion, perception is the most empirical thing we have. The fact that a group's perception converges means that, regardless of whether what they have converged upon is true or correct, they should be aware of the convergence. If it turns out to be useful information, congratulations. If it is merely a misinformed collective bias, that too is important for the group to be aware of.