Final Exam: Research Proposal for End-to-End Approach from Discovery to Hypotheses

This term, several of the papers I presented to the class involved the themes of idea generation across disjoint domains, literatures, or even groups of people. A common approach was to look for unknown similarities.

(Burt, 2004) identified the benefit of idea generation when looking across different clusters of people. (Kostoff et al., 2008) demonstrated the results of finding unknown connections between disjoint topics of literature and applied this technique to different subject domains. (Spangler & Wilkins, 2014) took this further by computationally deriving potential research hypothesis that were likely to produce outcomes from the overwhelming volume of literature. Indeed, Spangler et al. were dependent on having a specific problem to solve at the beginning of their project to develop their technique.

Burt, Kostoff, and Spangler all benefitted from the varying participation of subject matter experts to help guide, define and judge their results for the relevant subject domains. The identification of experts can therefore be assumed to be an essential ingredient for successful outcomes. If the experts have not already been included from the outset, there is a need to identify potential relevant experts not only for the evaluation of the literature analysis; but also who could possible turn the resultant discoveries into actionable research proposals and perform said research.

To that end, this term (An, Yu, & Li, 2014) presented research that identified institutions by their research domains and potential for collaboration. A post-presentation discussion with Lu An revealed colleagues who have been working on identifying people and research teams by their subject expertise.

I am proposing to bring the three together to develop a fully realized end-to-end path from discovery to hypotheses and identified researchers.

**Problem Definition**

Identifying hypotheses that are likely to produce outcomes is important as researchers seek to obtain funding for their projects. Being able to demonstrate some indicators of potential success would be helpful. Areas where there has been significant previous research lends credibility; however, convincing organizations to fund proposed research that has little to no reason for successful outcomes can be challenging.

**Current Practice**

There have been successes in identifying topic linkages between disjoint literatures. The Literature-Related Discovery and Innovation (LRDI) technique has shown promising
discoveries by mining disjoint literature and identifying previously unknown linkages (Kostoff et al., 2008).

Text mining has also been used to reduce an enormous number of potential solutions to a problem to a smaller number that are more likely to be a solution and thus warrant future research (Spangler & Wilkins, 2014).

Kostoff’s LRDI technique allows for discovery whereas Spangler’s approach starts with a problem from which it derives potential hypotheses for further research. However, there is no clear path from discovery of possible connections in the literature to actionable research.

A challenge with computational knowledge discovery is having the subject matter expertise to define the problem, query and evaluate the results. In some cases, an expert involved from the project from the outset provided the guidance and judgment to validate the connections resulting from the discovery process.

As we’ve seen in Spangler and Kostoff, one thing that made the results of LRDI and Automatic Hypothesis Generation more successful is the involvement of experts to guide the development of subject characterization, formulation of the queries and interpreting the results. Many of these partnerships were formed at the beginning of the research study; however, there are cases where the appropriate expert is not known and could be the link between taking the ideas resulting from LRDI and generating actionable research proposals. (Ling, Kang, Johns, Walls, & Bindoff, 2008) recognized this problem and proposed a hybrid approach of involving subject matter experts throughout process.

Identifying experts by authorship of the retrieved literature does seem like a solution; however, it is difficult to know which authors have the relevant expertise and possible that the experts exists outside of the authorship network. (Liu, Ye, & Liu, 2008) found that a single source of expertise identification is often not sufficient and several sources need to be merged together. (Liu et al., 2008) describes building a semantic repository of experts and others have done similar research in connecting people with research topics.

**Proposed Approach**

The proposed approach is comprised three identifiable phases:

1. Literature Related Discovery and Innovation,
2. Expert Identification, and
3. Automatic Hypothesis Generation.

These phases are not necessarily performed exclusive to each other depending on the complexity of the problem domain. In fact, this proposed approach recommends beginning expert identification at the same time as identifying the problem and subject domains.

(Ling et al., 2008) did recognize that different domains and data sources may require different levels of expert involvement earlier in the research depending on the complexity. Initially, this may result in redundancy of performing an initial LRDI to obtain subjects and terms to use in identifying experts who will then refine the subsequent, more intense LRDI activity.
These experts will then turn the discoveries into suitable problems and queries for Automatic Hypothesis Generation and validate the resulting hypotheses. Ideally, the same experts would be able to continue on with creating the research proposals and performing the research. However, those experts may be more suited for identifying additional specialists for that activity.

Impact

The expectation of this approach is for the researcher population to experience faster throughput of potential research topics that are “shovel ready” by:

- Reducing time spent in preliminary review and analysis of the literature and
- Identifying potential areas for more in-depth investigation.

Which should:

- Result in novel discoveries,
- Produce actionable hypotheses,
- Shorten the time to proposal, and
- Demonstrate a higher probability of success to funding organizations.

Challenges

In a conversation with Lu An, she pointed out that a major challenge in expert identification and classification remains the disambiguation of names. Names are not necessarily consistent across different data sources and languages and that has been a challenge in other researchers making similar attempts.

Research Plan

This research will require investigating the developments of LRDI since (Kostoff et al., 2008) and select LRDI techniques that are not only successful; but assist with identifying subject terms usable in Expert Identification.

At this point, the least defined portion of this approach is evaluating the different approaches for identifying experts by research domains and specialties. As (Liu et al., 2008) found that multiple data sources are necessary to identify experts, combining several techniques such as authorship, citations, social network analysis (Chou, Yang, & Lee, 2007; Fu, Xiang, Liu, Zhang, & Ma, 2007), and self identification may be what proves to be effective. Both (Chou et al., 2007; Liu, Liu, & Liu, 2007) remind us the importance of ranking not only the resulting experts as relevant to the topics; but also the importance of the data used in the analysis. Likely, investigating these techniques will require a significant portion of research time.

Finally, Automatic Hypothesis Generation (Spangler & Wilkins, 2014) is relatively recent and will need to be assessed for scalability for relationships other than the degree of similarity that was the basis for the algorithm.
To develop this process, several test research topics will need to be identified including, but not limited to medical/healthcare, social science, engineering, applied science, etc. Experts from the different topics chosen will be needed to validate the results along the way.

**Evaluation**

This approach can be evaluated in several ways. One is to measure rate of research proposals created as a result of this process and then the rate that receive funding. Another, longer-term measurement would be to track research funding and authorship directly emerging from this research activity. And finally, the results of each test subject should be thoroughly evaluated by experts as to how its contribution to their respective fields.

**References**


