REMatch: Research Expert Matching System

Md Iqbal Hossain¹, Stephen Kobourov¹, Helen Purchase², Mihai Surdeanu¹
{hossain,kobourov,msurdeanu}@cs.arizona.edu, helen.purchase@glasgow.ac.uk
¹Department of Computer Science, University of Arizona, USA
²School of Computing Science, University of Glasgow, Scotland

Abstract—We describe a system designed to process, analyze and visualize academic data, from research papers and research proposals to lists of courses taught, consulting, internal and external service. This can be helpful in identifying experts in a given field for future collaborations, as well as in putting together strong multi-disciplinary teams to apply for future research funding. Our REMatch system aims to support such tasks by leveraging natural language processing, machine learning, and interactive visualization. Specifically, REMatch provides a functional system that implements in-the-browser, map-based interactive navigation of a large underlying network, supporting semantic zooming, panning, searching, and map overlays. A prototype of the system is evaluated with a small-scale case study.

I. INTRODUCTION

The explosion of data leads to unexpected effects, where solutions to critical problems are overlooked, since many tasks need to cross several disciplines or domains that produce considerable amounts of data but interact only minimally. Swanson predicted this problem more than three decades ago, and described it as undiscovered public knowledge:

“Knowledge can be public, yet undiscovered, if independently created fragments are logically related but never retrieved, brought together, and interpreted.” [46].

In this work we focus on the complex task of matching academic calls to proposals. This task impacts universities and other research-oriented organizations, which must constantly answer difficult questions such as: How can we identify experts in a given field? How can we identify gaps in our strategic areas of expertise? How can we match calls for proposals with the right set of experts? How can we put together a strong multi-disciplinary team to apply for big research proposals?
Importantly, although these are complex tasks, there is great deal of data from sources such as research papers, research proposals, and courses taught that can be used to address them. This information can also be used to identify current strengths and weaknesses, as well as finding patterns and trends over time. The main challenge and our major contributions are in gathering and processing the needed data, putting together a collection of new and existing tools, designing an intuitive interactive interface, and packaging all of this in a functional system that non-experts can use.

A. Context

Encouraging research experts (tenure-track research faculty and research scientists) to apply for external funding is an important task for university research offices. This task is simultaneously very difficult (given thousands of research experts across dozens of colleges and hundreds of departments), while also of critical importance (given the major role of research funding in the face of steady cuts in state and nationwide education funding). In the rest of this paper we focus on the University of Arizona (UA), although the design principles can be applied to other universities.

The Research Office (RO) staff collect information about calls for funding from a wide range of sources (e.g., federal funds, industry proposals, and named foundations), with the aim of alerting research experts for whom the call is relevant. In particular, they would like to encourage people who might not have considered the call as being one for which they should apply, or who might have missed the call altogether.

The continuous stream of information (calls for funding proposals, information requests, white papers) is monitored and filtered by the RO staff, and a weekly digest is emailed to all university research experts. Small specialist teams (e.g., physical sciences, arts and humanities, clinical and biomedical) consider the calls in their particular area, and, after filtering, do one or both of two things: (a) they forward the calls to the relevant College Dean of Research (e.g., Engineering, Arts and Humanities, Science) asking them to in turn forward to appropriate faculty in their colleges, and/or (b) they directly contact particular individuals who are known to have specialist knowledge in the area relating to the call. When contacting individuals directly, RO staff rely on internal information (faculty profiles) and commercial tools (e.g., the PIVOT database, https://pivot.cos.com/). Despite these resources, RO staff readily acknowledge that a major source of knowledge about the research expertise of the institution’s researchers is “in their heads” – personal knowledge that they have built up over time through networking events and individual contacts, and which is lost when RO staff members change jobs.

It is the process of “finding the perfect fit” that the RO staff welcome support with – knowing exactly who would be the best people to target and encourage to respond to a particular call for funding. They would like to reduce the amount of targeting emails that they send (so as to cut down on emails considered “spam” by the university research experts), while being confident that the sources of information that they use in this targeting are robust enough to ensure both high precision (the people targeted are appropriate) and high recall (appropriate people are not omitted).

B. The REMatch System

The Research Expert Matching (REMatch) System relies on data gathered from university databases (e.g., current staff, research proposals), as well as external sources (e.g., research publications, research funding awards). This data is analyzed using machine learning (ML) and natural language processing (NLP) components and visualized with a interactive map-based network and overlays. In particular, given a specific call for proposal (CFP) or a research paper (or even a bit of relevant text from a CFP or paper), we can find the research experts who best match the query and locate them on the map. The query can be refined by specifying one or more colleges (e.g., show only experts in the College of Science), and further refined by marking some of the phrases and words from the input as “stop-words” (i.e., contentless words). Each person matched is associated with an explanation, showing which topics and terms were matched, and therefore explaining why that person has been included in the list. Exploring the map can show additional information, such as past collaborations, past funding data, and citation counts.

We gather data from different sources (several university databases, as well as external ones such as Google Scholar) to build the collaboration network. We then use multi-level force-directed placement, node overlap removal, and clustering algorithms to represent the network as a map. Nodes, node labels, polygon colors, and edges are transformed into Google Map objects, which are then drawn in the browser using the Google Maps API. Seven different level-of-detail (semantic zoom levels) are precomputed, which entails determining which nodes are present on a given level, computing label font sizes, and ensuring no label overlaps. To process a text query and find research experts that match it, we extract research topics and terms associated with the query. Then we perform multiple information retrieval queries on a precomputed Lucene index, which includes information about each expert’s research topics, publication details, and grant proposals. Figure. 2 shows an overview of our system.

The REMatch system provides several novel features: (a) leveraging knowledge and information from different domains to find relevant research expertise and research experts; (b) combining machine learning, natural language processing and visualization techniques in one functional system that (unlike other stand-alone graph-drawing tools) is implemented in the browser, and is therefore readily accessible; (c) providing interactive map-based visualization of the network of research experts and supporting common map-exploration interactions such as multiple zooming levels, semantic zooming, panning, and searching; and (e) overlaying additional information such as collaborations, citations, and funding levels. The current prototype of REMatch is already used by senior managers at the university to identify potential new research collaborations and to create multi-disciplinary research teams. As an open-
source system, we expect that REMatch can be customized to different institutions and that additional features will make it useful beyond the scope of the current prototype.

II. RELATED WORK

There is a great deal of related work in different domains: from collaboration networks and topic analysis, to visualization techniques for text data and large graphs.

Research collaboration networks are widely studied [24], [33]. Words from paper titles have been used as indicators for the content of a research topic, which in turn are used to create topic visualizations [17], [47], [56]. Some earlier approaches focus on analyzing specific journals, conferences, or research areas, e.g., analyzing computer science conferences and journals [17], trends in computer science research [15], the International Conference on Data Mining (ICDM) [30], publications in data visualization [21], Domenico et al. [11] quantify attractive topics (i.e., topics that attract researchers from different areas). Sun et al. [44] build a network, with computer science conferences as nodes and edges between two conferences with common authors. Map-based visualization has been used for document visualization [22], [42], [53]. Dunne et al. [13] built a tool to combine reference management, citation text, automatic summarization, ranking and filtering, and network visualization for documents.

Citations are considered an important contribution measurement [55] and are used in visualizations of scholar profiles [35] and paper recommendation systems [51]. Citation data from the Web of Science [34] and from Microsoft’s Academic Graph [25] have been analyzed. CiteRivers [21] and Cite-VIS [43] analyze and visualize IEEE VIS conference citations, as do Ke et al. [26].

Also related to our work are many graph visualization techniques and tools; see surveys by Von Landesberger et al. [49] and Vehlow et al. [48]. Graph layout algorithms are provided in several libraries, such as GraphViz [3], OGDF [10], MSAGL [32], and VTK [40], which however, do not support interaction, navigation, and data manipulation. Visualization toolkits such as Prefuse [20], Tulip [7], Gephi [8], and yEd [52] support visual graph manipulation, and while they can handle large graphs, their rendering does not: even for graphs with a few thousand vertices, the amount of information rendered statically on the screen makes the visualization difficult to use. Further, there are research papers that describe interactive multi-level interfaces for exploring large graphs such as ASK-GraphView [6], topological fisheye views [18], and Grokker [37]. Software applications such as Pajek [12] for social networks, and Cytoscape [41] for biological data provide limited support for multi-level network visualization. These approaches rely on meta-graphs, made out of meta-vertices and meta-edges, which make interactions such as semantic zooming, searching, and navigation counter-intuitive. Not many of the tools and systems above provide browser-level navigation and interaction for large graphs.

Liu et al. collect publications and projects of potential supervisors from internal data sources of University of Leeds so that new PhD applicants can find potential supervisors [27]. As this approach relies on manually defined rules, it does not easily extend to larger scale datasets such as ours. Automated recommender systems such as blended recommending [28], expertise recommender [29], ruled-based mapping [31], and group recommender systems [50] deal with large-scale problems such as ours, but are not clearly applicable to academic publications, research grant proposals, and course descriptions.

Commercial organizations provide access to (and, in most cases, visualizations of) data about research activity for an institutional fee. SciVal (Elsevier, www.elsevier.com/solutions/scival) “offers quick, easy access to the research performance of 8,500 research institutions and 220 nations worldwide,” and Academic Analytics (www.academicanalytics.com) focusses on research universities in the United States and the United Kingdom, specifically supporting “the strategic decision-making process as well as a method for benchmarking in comparison to other institutions.” In profiling an institution, Pure (Elsevier, www.elsevier.com/solutions/pure) “aggregates your organization’s research information ... enables your organization to build reports, carry out performance assessments, manage researcher profiles, enable research networking and expertise,” while InCites (Thomson Reuters, http://clarivate.libguides.com/incites) allows you to “analyze institutional productivity, monitor collaboration activity, identify influential researchers, showcase strengths, and discover areas of opportunity.” Universities pay hundreds of thousands of dollars for these services, typically in the form of multi-year contracts, yet these services lack some critical
pieces of information that we do have access to (e.g., past research proposals) and do not match research experts to calls for proposals.

Our work builds on this previous knowledge, and proposes an integrated search and visualization platform for the important task of matching research experts with calls for proposals, a challenging task that is currently handled manually by university staff.

III. REMatch System Overview

The REMatch system relies on a wide range of data collected from the following internal and external sources:
1) UA Vitae: an internal online system for university faculty and staff, which includes research publications;
2) UAIR: university internal analytics and institutional research, which keeps track of staff arrivals and departures;
3) UAR: UA’s research activity system, which keeps track of research proposals and awards information internally;
4) Google Scholar Profiles: we scraped UA Google Scholar profiles for publications data (e.g., title, abstract, citations, research areas);
5) Google Scholar Topics: we scraped half a million research profiles from the top one thousand universities (according to the Center for World University Rankings [2]), in order to build a topic network connecting co-occurring research topics.

These data sources are integrated so as to associate a rich individual research profile with each member of the institution.

The system interaction entails the entry of a URL address for a call for proposals (CFP) or plain text, typically the content-rich passages from a CFP. The output is a ranked list of research experts who match the CFP, i.e., a list of people who RO staff might reasonably contact and encourage to respond to the CFP. Clicking beside each name reveals the relevant matching research topics and terms. A bar chart shows the number of matches in each college and clicking on it filters the results. Another bar chart shows the number of matches by type of position (e.g., Assistant Professor) and clicking on it further filters the results. The set of the keywords extracted from the CFP is shown - clicking on any of them makes them stop-words for this query (removing them from the set) and the list of experts is updated with this reduced set. The collaboration map highlights the research experts in the current matched list - this map can be navigated and explored to find more information about the highlighted experts, including their collaborators, the nature of these collaborations (titles of joint papers or research proposals), level of current funding, and number of citations.

The steps in this process are therefore (1) building the university-wide collaboration map, (2) building a database of research expertise, (3) extracting research topics and terms from the CFP, (4) identifying the relevant faculty members, (5) displaying them on the map, (6) filtering with respect to college, (7) removing irrelevant keywords. We discuss each of these steps below.

IV. Research Collaboration Map

We chose to use a map-based visualization to show the results in the context of cross-university collaboration. Maps have been shown to be effective and memorable representations of networks [19], [38], [39]. In our case, the underlying data is a collaboration network, in which each node is a person, and an edge exists between two nodes if the corresponding pair have past collaborations (e.g., joint research paper, joint research proposal). Collaborations are based on publication data in the databases scraped (described in the previous section). Named-entity recognition is one of the challenging issues when working with multiple datasets. In all internal university systems, researchers are identified by an unique ID. Each Google Scholar profile also assigns unique IDs. We match IDs from different databases using a semi-automated system based on OpenRefine [5].

We use the GMap framework [19] to generate a map-like visualization of the collaboration network and extend it to support semantic zooming. The process can be summarized as follows: (1) embed the collaboration network in the plane, (2) group nodes into clusters, (3) create the geographic map representation, (4) compute multiple level-of-detail views, and (5) provide support for interactions (pan, zoom, search) and overlays.

We embed the network using a scalable force-directed algorithm (sfdp from graphviz) and then group the nodes using k-means clustering. To create the geographic-map look, we use a modified Voronoi diagram based on the obtained embedding and clustering. The geographic regions are colored such that no two adjacent countries have colors that are too similar, using the spectral vertex labeling method [19]. We use the GraphViz implementation of node-overlap removal provided by prism. Note that prism provides non-overlapping labels only for the complete basemap (showing all nodes), and not for the other 6 level-of-detail views, needed for semantic zooming.

The semantic zoom in REMatch requires modifications to nodes, edges, clusters, and heatmap overlays; see Fig. 3. We use the Google Maps API which handles most of these issues, with the notable exception of node-overlap (and hence node-label overlap), which is a natural side effect of zooming. To ensure that neither nodes nor labels overlap on any level of detail, we compute different node visibilities for different zoom-levels. For each level, we sort the nodes by their weight (node-weight is determined by the degree of the node, or amount of funding, or number of citations). We make the i-th node visible on the j-th level if the bounding box of the i-th node does not overlap with the bounding boxes of nodes 1, 2, · · · , (i−1). This algorithm requires O(n^2) time but it could be improved using different techniques as in [14].

V. Matching Research Experts

In this section we describe how we find the research experts who best match a given text query, (e.g., a CFP). We treat this as an information retrieval (IR) problem, where the text query forms the (large) input query, and each individual researcher
is represented as a single “document” containing all associated research materials (e.g., research topics, publications, submitted proposals). We further extend this process with a ML component, as described below.

### A. Building a Database of Research Expertise

We first assign research topics to all researcher experts. In general, Google Scholar (GS) profiles provide self-reported description of a researcher’s interests or research topics. However, many researchers do not have a GS profile and many have a profile that is very sparse (containing only one or two research topics), which complicates the IR task addressed in this section. To mitigate this problem, we implement an ML component that predicts five research topics for such researchers.

Out of the current total 2,187 researchers at UA, 1,127 have GS profiles. Of these, 314 provide the maximum possible five research topics and no prediction is necessary. For all others we use the method described below. When a researcher has no GS profile we use the top 5 predicted topics. When a researcher has a GS profile with fewer than 5 topics, we augment the given topics with our prediction to again obtain exactly 5 topics. The prediction process relies on a multi-class, multi-label classification with regularized logistic regression which is implemented in the LiblineaR package.

**Task:** Determine Top-5 research topics from an expert’s data.

---

**Data:** We scraped GS profiles, gathering the self-reported areas of expertise and publication details. Note that multiple research topics are typically mentioned in each profile. We consider all publications in a given profile as belonging to each of the research areas provided in the profile. We use this data as training data. To predict the Top-5 research topics associated with each researcher, we use all associated data (research publications, grant proposals, etc.).

**Data preprocessing:** To prepare the dataset for the machine learning model we first extract important phrases (i.e., noun phrases) from the documents. We rely on the Natural Language Toolkit (NLTK) [4] for this task, using the grammar $\text{NP} : \{(<\text{JJR}> | <\text{JJS}> | <\text{JJ}> | <\text{NN}> | <\text{NNS}> | <\text{NNP}> | <\text{NNPS}> \})^*$. We then stem the phrases with using a series of operations from Lucene [1]: StandardFilter,LowerCaseFilter, StopFilter, and SnowballFilter. Finally we compute term frequency as shown in Fig 4.

**Training:** Formally, we consider the self-reported research topics in GS profiles for the 1,127 academics as labels ($Y$). For all researchers, we build a matrix $X$ from their publications: each row corresponds to an individual publication; each column represents a noun phrase found in the corresponding text, with a value set to its term frequency (TF) in the corresponding document. The label(s) assigned to rows in $X$ are taken from all research topics reported by the corresponding researchers in GS. The matrix $X$ used for training contains 38,886 rows and 8,462 columns. This matrix is used to train the multi-class logistic regression classifier.

**Prediction:** At prediction time, the model returns classification probabilities as a $n \times k$ matrix, where $k$ is the number of research topics, and $n$ is the number of researchers to be classified. We take only the top five columns for each person, i.e., the top predicted research topics of a person. Where needed (e.g., for researchers with fewer than 5 topics in their GS profile) we augment the list to obtain the Top-5, making...
sure there are no repeats.

B. Extracting Google Research Topics

In a topic network, topics are the nodes and edges indicate that topics are related to each other. Extracting topics from research articles (with topic co-occurrence within an article indicating topic relationship) is a popular approach to creating a knowledge network [17], [54]), but these methods do not allow for easy identification of general topics (e.g., mathematics, physics), nor do they include very specific topics (e.g., symmetry detection algorithms, interactive graph visualization) as nodes.

Our approach for creating a topic network relies on the assumption that people know the topics that they work on: nobody is better placed to categorize researchers’ topic areas than the researchers themselves, and, while document analysis might automatically identify and extract topic labels from an article, only the researchers who wrote the article know precisely the key topics of the paper. We therefore use the self-reported areas of study in the Google Scholar (GS) database, as defined by researchers themselves and the co-occurrence of topics within a researcher’s list indicates a relationship between them in our knowledge network. Specifically, the topics network is generated using the following steps:

1) Data Scraping: Our data gathering is limited to GS entries associated with the world’s top 1,000 universities (according to the Center for World University Rankings [2]). We extracted the institution IDs from GS and then scraped the URL associated with each institution to collect research profiles of all individuals associated with the institution, focusing on the list of research topics from each research profile. The total number of topics extracted from this raw data was 190,137, but after standardizing the topic separators within the topic list, and using beautifulsoup [2] to tidy up html tags for consistency, the number of distinct topics rose to 222,459.

2) Data Cleaning: We removed leading or trailing spaces, unnecessary punctuation and control characters, and duplicate topics. Many topics were phrases or composite terms (e.g., “statistics for neuroscience,” “data and model management,” “group theory and combinatorics,” “symmetries of graphs”); we removed conjunctions (and, or) and other words with no semantic weight (for, of), thus splitting topic phrases into their constituents.

3) Topic Correction: Typos and acronyms frequently occur. We used Google’s OpenRefine [5] to identify and resolve typing errors, and to find alternate representations of the same topic [9], [16], [23] (e.g. “Computer Human-Interaction” is equivalent to “Human-Computer Interaction”; “Primary education” is the same as “Elementary education”). This process reduced the number of unique topics to 210,588.

4) Topic Removal: We dropped topics that were associated with four or fewer people (aware that these topics might be topic labels in which there were typing errors that were not captured by OpenRefine), and topics that we identified as not being in English. This reduced the number of topics to 39,067.

5) Topic Merging: Merging was required for topics that are similar, but are listed slightly differently; for example, “algorithm,” “algorithms,” “algorithmics” are all the same topic, as are “organization,” “organizational.” We used snowball [36] to find the root word by applying stemming processes (removing endings such as -s, -ed, -ing). “Algorithm,” “algorithms”, and “algorithmics” thus all become “algorithm;” however “applied” and “applications” become the meaningless term “appli.” To mitigate against nonsensical resolution, we choose the main topic to be the one with the highest frequency amongst all topics with the same stem.

C. Lucene Indexing

As mentioned earlier, we treat the problem of matching academics with a given research topic as an IR task, which we implemented using Apache Lucene [1]. In particular, we construct a Lucene document for each individual researcher, with two data fields. The first field (F1) contains titles and abstracts of research papers and grant proposals. The second field (F2) contains the research topics of the corresponding person (from their GS profile, or predicted as described above). The third field (F3) contains information from biographies/CVs, list of courses taught, internal and external service, consulting, etc.; see Table I.

This approach allows us to combine three sources of information in a single query: F1 contains data that is (more or less) objective since it comes from academic publications, but it may be too verbose, F2 contains precise information, but which may be biased since it is self-reported, and F3 contains high level research areas that are not covered by F1 and F2.

D. Query Pre-Processing

Given an input query (e.g., a URL for a CFP, plain text, or a collection of keywords), we convert it to an actual Lucene query as follows: The text is cleaned, tokenized, and stemmed. We next compute the term frequency of unigrams, bigrams, and trigrams, which become our candidate research topics. Then we match the candidate research topics with the research topics extracted from GS, by taking the intersection of the two sets of topics (in our network we only include topics that

Table I: Data summary.

<table>
<thead>
<tr>
<th>Field</th>
<th>Data Description</th>
<th>#Records</th>
</tr>
</thead>
<tbody>
<tr>
<td>F1</td>
<td>Research paper titles, abstracts</td>
<td>165501</td>
</tr>
<tr>
<td></td>
<td>Grant proposals</td>
<td>44970</td>
</tr>
<tr>
<td></td>
<td>Abstracts of grant proposals</td>
<td>1325</td>
</tr>
<tr>
<td>F2</td>
<td>Top 5 research predicted areas</td>
<td>2554</td>
</tr>
<tr>
<td></td>
<td>Courses taught</td>
<td>240823</td>
</tr>
<tr>
<td></td>
<td>Internal and external service</td>
<td>18256</td>
</tr>
<tr>
<td></td>
<td>Teaching and research interests</td>
<td>1706</td>
</tr>
<tr>
<td></td>
<td>Consulting</td>
<td>1690</td>
</tr>
<tr>
<td></td>
<td>Biography</td>
<td>1222</td>
</tr>
<tr>
<td></td>
<td>Research activity / Work in progress</td>
<td>2032</td>
</tr>
<tr>
<td>F3</td>
<td>Research goals and progress</td>
<td>6844</td>
</tr>
</tbody>
</table>
occur more than once in the 500,000 analyzed GS profiles). The result is a collection of known research topics associated with the input query. This set of research topics is denoted by $Q_1$, while the list of unigrams in the input query is denoted by $Q_2$.

**E. Lucene Queries and Score Merge**

Using this information, we construct three separate queries, and compute a combined score through linear interpolation (with hyper parameters $\lambda_i$), as follows:

$Score_1$: Normalized score for phrase query on index field $F_2$, using query $Q_1$;

$Score_2$: Normalized score for query on index field $F_2$, with query $Q_2$;

$Score_3$: Normalized score for query on index field $F_1$, with query $Q_2$;

$Score_4$: Normalized score for phrase query on index field $F_3$, using query $Q_1$;

The overall score $S$ for a given research expert $p$ is then computed as: $S(p) = \sum_i \{\lambda_i \times Score_i(p)\}$ where $\lambda_1 + \lambda_2 + \lambda_3 + \lambda_4 = 1$. Since $Score_1$ captures a match between the research areas of the input query and the expert’s research areas $\lambda_1$ is larger than $\lambda_2$ and $\lambda_3$. In our implementation we use $\lambda = [0.3, 0.1, 0.1, 0.5]$, that is, $F_1$ and $F_2$ contribute a combined 50% and $F_3$ the other 50% of the total score.

**VI. Map Overlays**

Currently the REMatch system provides the following overlays:

- **Citation Heatmap**: highlights highly-cited experts, based on citation counts or normalized citation counts
- **Funds Overlay**: shows funding levels (overall or from specific funding agencies) using a green circle with radius proportional to current funds associated with each research expert
- **Edges**: shows all connections between research experts, as by default these are not shown, relying on Tobler’s first law of geography: “everything is related to everything else, but near things are more related than distant things”
- **College Overlay**: highlights all research expert (nodes) from a given college and/or all collaborations between pairs of research experts from the same college
- **Individual Details**: We provide basic search functionality which locates people on the map. Clicking on a node shows the name of the person, along with the number of connections, citations count, and college affiliation; edges to collaborators are also shown; see Fig. 5. By clicking on an edge we can see information about the collaborative work.
- **Label Size**: Depending on what we want to show, we can change node importance (which is reflected in the font sizes of the labels) by considering the degree of the node, the research funding amount, or number of citations associated with this node; see Fig. 6.

Several of these overlays can be shown concurrently, which helps visualize different data dimensions, e.g., identifying research experts who are producing highly cited work, research experts who are well-connected in the university, and research experts who are good at securing funding; see Fig. 5.

**VII. Implementation**

Given a CFP, a research paper, or even a bit of relevant text from a CFP or paper, our system matches the best experts (as described in Section V) and shows the results on the map. We can filter by college and explore the explanations for the matches. Moreover, the system shows the important single terms identified from the query, any of which can be marked as stop-words. Clicking on a word removes it from the query...
Figure 6: Node and label sizes are determined by one of three variables: number of collaborations (node degree), funding amounts, and citations. This is illustrated for the same portion on the map above. Note that some nodes disappear (when they drop below the threshold size for the current level of detail) and that their font sizes change.

and can improve the results. Figure 1 shows a screenshot of the system.

We use a variety of tools to clean, store, and process our data: mongodb scripts, sqlite, python, R, Java-Lucene, openrefine. Google maps API and jquery are used for map drawing and to handle interactions in the web application. We run python-django for the webserver and mongodb for database storage and queries.

GMap produces a “basemap” from the given graph, which is a static image that is not well-suited for user interaction, such as zooming, panning, and searching. We enable interactions with the help of the google maps API [45]. Specifically, we take the output from GMap and convert it into google map objects, i.e., google.maps.SymbolPath, google.maps.Polygon, google.maps.Polyline, etc. For the web interface we provide 7 levels of details, showing different subgraphs, depending on the zoom level.

VIII. Evaluation

We performed a case-study evaluation with two Research Office (RO) staff members, both of whom perform the task of matching funding calls to faculty on a regular basis – one in the Medical and Biological Sciences, and one in Environmental Science. They each brought two relevant funding calls to the meeting, together with a list of people they would target for these calls, having followed their normal processes in devising these lists. We gave a live demonstration of the system – for each call, we entered relevant content-rich paragraphs into the system, showed the list of names, filtered by colleges as appropriate, and discussed the results with them. There were three responses to each of the four lists:

1) Why is X there? This happened in the cases where the RO staff members know a specific person well, and are surprised to see them matched for the particular call.
2) Why is X not there? This happened for people included on the RO staff member’s list who are ranked low on the REMatch list or do not appear at all.
3) I don’t know X. These are people that the RO staff member does not know at all, and would be interested in finding out why they were included in the list.

(1) and (2) were responses that are a result of personal knowledge of the RO staff – they rely on the information about faculty that they keep in their heads – information that can be lost with staff turnover and which is hard to collect given thousands of university researchers. (3) is a positive response, since this encourages the RO staff to find out more about university researchers who they do not already know. It also is useful for new RO staff members.

In discussion, we discovered that anomalies were closely related to context. For example, for one of the calls, an extensive multi-disciplinary team is needed, spanning several colleges – requiring separate filtering for each relevant college. In another, people were missed because they have joint cross-college appointments (and did not show up in the first college selected). For one proposal, a senior team of established researchers was needed, and so names of junior university researchers are not good matches. We discovered that several people were included on the REMatch list because of very generic terms (e.g., “environment”) and the facility to “turn off” such all-encompassing stop-words was welcomed.

Overall, the prototype REMatch system received a very positive response from the RO staff, despite the few anomalies in these four specific calls. For those names that they had not thought of (category (3) above), they were intrigued (“[…] is an interesting fit. Someone like that would be good.”; “none of these are totally wacky”) and recognized the usefulness of being directed to people they had not previously considered. One approach that they use that is currently not supported is the ability to identify people who have previously failed in a funding bid. This information would enable the RO staff to focus on people who might need support in further applications, or to identify other appropriate opportunities that a failed bid might be recycled for. However, they acknowledge that
this is sensitive information. From a process perspective, the two RO staff members were impressed with the fact that the text entered can be customized – other research management systems (for example, Academic Analytics) require the entry of specified keywords, making them “not as intuitive and easy to use as this system.” In effect, our system serves as a pre-processing step by extracting the relevant keywords.

The RO staff members were confident that the REMatch would be a very useful complement to their current processes, and would save them a great deal of time. The ability to “easily refine and hone down” results means that they have control over the search, while taking advantage of the explanations, collaboration map and faculty profile information that is easily accessible.

IX. DISCUSSION AND LIMITATIONS

Anecdotally, our REMatch system seems to be fulfilling its purpose well. For example, it placed all 8 of the actual principal investigators and senior personnel of a recent proposal sent to the National Science Foundation TRIPods program among the top 15 matches (as shown in the accompanying video, available at https://uamap-dev.arl.arizona.edu). Note that this particular CFP is recent and was not included in the training data for REMatch. Similarly, out of four recently submitted proposals (that were also not part of the training data), REMatch identified three of the Principal Investigators who submitted proposals to these programs. The University’s Research, Discovery & Innovation Office is using the REMatch system to put together research teams for large-scale, multi-disciplinary projects, even though it is still under development. We plan to add new features based on suggestions we have already received.

We rely on several internal and external sources for gathering the data that makes REMatch work. For example, relying on Google Scholar has some advantages (e.g., a large amount of information) but also many disadvantages (e.g., the data is not curated). Further, different research areas differ in the extent of their representation in Google Scholar. For example, there seem to be many more computer science and physics profiles than history and psychology ones. Data from both internal and external sources needs to be updated regularly in order to provide current and relevant results.

Before arriving at our four-level scoring function (described in Section V-E) we tried several simpler approaches, all of which resulted in poor matches. We considered treating both research experts and CFPs as documents represented by high dimensional vectors of unigrams weighted by frequency or TF/IDF (term frequency/inverse document frequency) and finding the best matches via cosine similarity between the angles of a given CFP vector and the research expert vectors. This resulted in bad matches due to the high importance of adjectives and articles. Limiting to noun-phrases did not help as common CFP terms such as grants, proposers, projects carried little signal. Treating such phrases as stop-words and even using a different English language corpus as a baseline did not yield the desired results. Several other attempts failed before we arrived at the four-level scoring function we currently use. Still, we hope to further improve our scoring function.

We are aware that our focus on the underlying computational processes and the map visualization means that our interface could do with some improvements with respect to human-computer interaction design, in particular when considering the nature of the users and their specific visual analytics tasks.

X. CONCLUSIONS

The REMatch system helps identify areas of expertise, as well as experts in a given field, which is useful in the context of putting together strong multi-disciplinary research teams. We consider the main challenge and our major contributions to be in gathering and processing the needed data, putting together a collection of new and existing tools, designing an intuitive interactive interface, and packaging all of this in a functional system that non-experts can use. In addition to the case studies above, we plan to formally validate several of its components, starting with the quality of the predicted research topics and the quality of the proposal-person matches.

Despite non-trivial limitations, REMatch is novel as it provides a functional system that implements in-the-browser, map-based interactive navigation of a large underlying network, supporting panning, zooming, and searching, as well as map overlays. The REMatch system is open source and a video showing the system in action is available at https://uamap-dev.arl.arizona.edu.

REFERENCES
