# On Using Disparate Scholarly Data to Identify Potential Members for Interdisciplinary Research Groups

Francisco Osuna
*Cyber-ShARE Center of Excellence*
*The University of Texas at El Paso*
*El Paso, USA*
*Email: fjosuna@utep.edu*

Monika Akbar
*Cyber-ShARE Center of Excellence*
*The University of Texas at El Paso*
*El Paso, USA*
*Email: makbar@utep.edu*

Ann Q. Gates
*Department of Computer Science*
*The University of Texas at El Paso*
*El Paso, USA*
*Email: agates@utep.edu*

***Abstract*—Supporting interdisciplinary research (IDR) requires detecting the expertise needed to solve complex problems and identifying researchers with that expertise. Universities have adopted various expertise systems, many of which use publications and keywords to identify experts. Research expertise is dynamic in nature as one's expertise may change over time. Relying solely on publications to infer research interests can be less effective in identifying potential collaborators as different types of scholarly activities demonstrate the change in research direction at different times. This paper uses disparate scholarly data to propose and evaluate different approaches for building research footprints and presents experimental results to show how these footprints perform in identifying potential members for IDR groups. Results indicate that grant data is a better predictor of IDR membership than publication data. The paper also describes two approaches for building IDR-specific classifier models, along with the accuracy of those models in identifying potential IDR group membership.**

***Keywords*-Scholarly data; interdisciplinary research; classification; research footprint.**

## I. Introduction

Today's complex scientific and social challenges require bringing in individuals who can contribute different perspectives, experiences, knowledge, and skills to advance education and research. There has been an increased emphasis on interdisciplinary research (IDR) and activities that support interactions needed to solve problems that cross disciplinary boundaries [1]–[3]. One challenge in supporting IDR groups is the ability to identify researchers who have similar or complementary expertise and knowledge to contribute to an initiative. Universities are adopting a number of different expertise systems, e.g., Vivo [4], Team Science Toolkit [5], to support sharing knowledge on expertise.

Many of such systems (e.g., Team science toolkit, Vivo) depend on user-provided keywords and data to identify expertise. It is important to note that researchers who select their own keywords may consider broader audiences of the expertise system, or those from their discipline. In the former case, keywords or concepts would more likely be abstract, while in the latter case, the keywords would be specific to the researcher's discipline. Other expertise systems deploy analytical approaches to extract keywords from publications. Platforms such as ArnetMiner uses publications to identify broad themes of research interests [6]. However, such rudimentary approaches often fail to identify different facets of research expertise. These systems also fail to make recommendations about which expertise can be beneficial to other disciplines.

This paper presents an approach to analyze scholarly data for identification of expertise at an institution by extending the research footprint of faculty members. A research footprint of a faculty member represents his/her research expertise and interests based on different scholarly activities. The proposed approach considers two types of scholarly data, publications and grant proposals, to build the research footprints. Experimentation is done using various approaches to generate the footprints, which are utilized later to identify potential membership in communities of practice (CoPs), i.e., groups of people with a shared domain of interest. The contribution of the paper is the study of approaches to generate research footprints from scholarly data through auto-extracted keywords, concepts related to the keywords, and full text comprised of the title and abstract of said scholarly documents. The experimental results presented in this paper demonstrate the impact of using different types of research footprints in detecting membership to interdisciplinary research groups.

The paper presents a review of similar work in Section II, followed by problem description in Section III. Section IV describes the approach used to generate research footprints associated with publications and proposal submissions. Later, generated footprints are used to identify possible matches with three CoPs. Section V presents the results of footprint generation and CoP matching. Section VI provides a discussion and future work.

## II. Related Work

One of the first steps of finding possible collaborators for an IDR group is to identify the research expertise of any given researcher. Another challenge is to identify the best match between IDR groups and a pool of researchers.

Research works have explored different approaches for building user profiles for various tasks. Diederich and Iofciu propose a Folksonomy-based approach for developing user profiles in order to detect communities of practice [7]. Their approach depends on the user's activity, specifically on the user-selected objects such as publications. Manual attributes of the objects, for example, author-selected keywords, are used to build tag-based profiles for the users. Others have proposed content-driven user profiling [8] [9], automatic tag recommendation for expertise profiling [10] using controlled vocabularies to capture expertise and generate research profiles [11], and social network, emails, and chat logs along with profile information for finding experts at the enterprise level [12]. User profiles in academic context are most often used for recommending scholarly articles or potential collaborators. Wang and Blei utilize both content and users' rating for making recommendations about scientific articles [13]. They use probabilistic topic modeling on other users' libraries for recommending unrated articles.

Researchers have used automatic keyword extraction and topic modeling to eliminate manual activities (e.g., tagging) for building user profiles. Probabilistic topic models, such as Latent Dirichlet Allocation, are used to detect latent thematic information from large text corpus [14]. Cross-domain Topic Modeling utilizes publications to recommend interdisciplinary collaborations while addressing challenges including sparse connection, complementary expertise, and topic skewness [15]. The Language Model extracts and ranks candidate key-phrases [16]. Automatic keyword extraction approaches, such as Maui [17], is able to determine main topics in documents by extracting keywords without using controlled vocabulary. Other approaches, such as graph-based keyword extraction methods, explore both the content of the document and the context [18].

## III. Problem Description

Research institutions and universities are focusing on building interdisciplinary research groups or communities of practice that bring together and support a diverse group of researchers who have an affinity for a particular research topic. Assembling such collaborative teams requires the ability to (i) identify the expertise needed for the collaborative research group or community of practice, and (ii) identify researchers with that expertise.

In academia, faculty members and researchers work on different scholarly activities related to their research interest and expertise. Such scholarly activities include writing grants, conducting research, advising students, publishing research outcomes, offering classes and workshops, and presenting research works. Each of these activities, when considered separately and as a whole, can provide insight into the expertise of the researcher. Research interest and expertise may change over time. Different types of scholarly activities demonstrate the change in research direction at different times.

The task addressed in this paper is to identify potential communities of practice (i.e., CoPs) for each researcher of an institution given the titles and abstracts of published articles and submitted grant proposals of the researchers. We use different types of scholarly activities — publishing research articles and submitting grant proposals — to build research footprints of each researcher. A *research footprint* is a set of words or concepts retrieved from the scholarly data of a faculty member that demonstrates his/her expertise. This footprint is used to identify potential membership of a researcher in CoPs.

## IV. Approach

Collaborative, in particular interdisciplinary, research has gained attention over the past few decades. The increase in collaborative research means researchers are more frequently contributing to different research projects. Thus, research expertise and interest of researchers are more likely to experience a shift over time. Research interests evolve based on different factors, including current challenges, demand, the mission of the institution, and funding status.

Expertise systems in academia are widely used to identify possible members for collaborative research. These systems host the profiles of researchers; most often the researchers identify a set of keywords to indicate their research interest. This practice leads to a collection of keywords that can be highly subjective. Other times, automatic keyword extraction approaches are used to extract keywords from researchers' publications. There are a number of challenges in this scenario, including selecting the best approach for building research profiles, detecting the research threads of a CoP, and identifying potential members with complementary research expertise for any given CoP.

The work presented in this paper partly addresses some of these challenges by presenting different ways of generating research footprints of researchers using different types of scholarly data. We investigate multiple approaches for generating footprints using keywords, concepts related to the keywords, and full text comprised of the title and abstract of scholarly data. The footprints are then used to detect alignment between a researcher with a given set of CoPs.

Figure 1 shows the approach presented in this paper. Data is collected from a number of different sources, some of which is used to build statistical models for each community-of-practice, and the rest of the data is used to generate research footprints for CoP members. Later, these footprints are used to identify any alignment with a given set of CoPs. CoP membership is used as the ground truth for assessing the results. The rest of this section describes each of these steps in details.

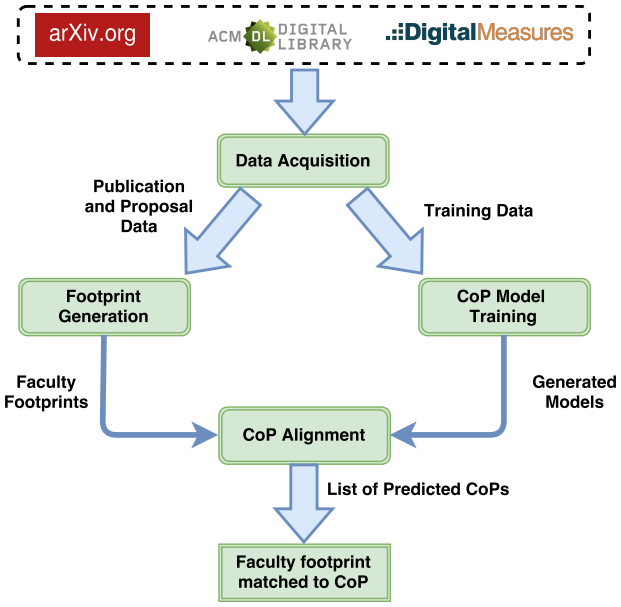


Figure 1. Generating research footprints and detecting alignments of those footprints with a set of CoPs.

### A. Data Acquisition

The proposed work utilizes different data sources in different stages of the process. Figure 2 shows some of the scholarly activities and possible data sources that capture information about these activities. The data sources can be broadly divided into two groups: internal data sources and external data sources. Internal data sources are comprised of repositories maintained by universities. For example, Digital Measures or Interfolio is used in many universities to capture information about the scholarly activities conducted by faculty members. Each university also has an office that manages grants and proposals.

There are a number of university-independent, external data repositories that collect or index information related to scholarly activities. For example, the National Science Foundation (NSF) has information on the title, abstract, and relevant metadata on funded projects. Scholarly works are often indexed by digital libraries, including ACM digital library, IEEE Xplore Digital library, PubMed, or Web of Science. These repositories are referred in this paper as external data sources. Many of the external, as well as internal, data sources provide APIs for collecting data. The work presented in this paper leverages partial ArXiv[1] data.

We retrieve each faculty member's publication data from Digital Measures and proposal data from the university's proposal submission system, which stores data about all the submitted, as well as funded, proposals. Publications and proposals, containing the title and abstract, are considered as complete and used for further processing. Thus, any publication or proposal without an abstract is discarded

[1]https://arxiv.org/

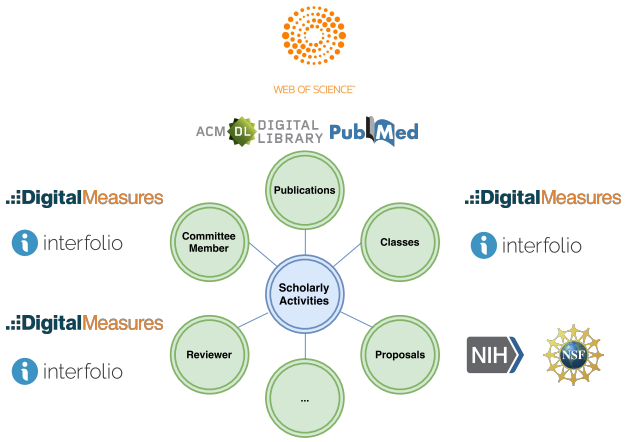


Figure 2. Data Sources for different types of scholarly activities.

based on being incomplete. This study does not make a distinction between the types of publications (e.g., journal, conference, workshop), as the intent of the effort is to identify expertise and *interest* in a CoP. No weight is placed on the status of the proposal, i.e., whether it is pending, funded, or not funded. All publications and proposals receive equal weight. Titles and abstracts from publications and proposals are used for the analyses. Once acquired, data is cleaned using standard data pre-processing steps, including parsing and stop word removal. The rest of the paper uses "grants" and "proposals" interchangeably, thus "grants" may not necessarily be awarded funding.

The work presented in this paper considers three existing CoPs at the University of Texas at El Paso: Cyber-Security, Smart Cities, and Neuroscience. These CoPs consist of a total of 47 members at the time of the study. Membership to these CoPs is open to any faculty member within the university. A member can choose to be part of multiple CoPs. The Cyber-Security is the largest CoP with 28 researchers. With 10 researchers, Smart Cities is the smallest of these three CoPs. Of the 47 researchers, five are members of multiple CoPs. Table I shows each CoP with the number of members, the number of complete publications (e.g., with abstract), and the number of complete grants associated with the members.

### B. Research Footprint Generation

In this paper, *research footprint* refers to a representation of research expertise as observed in the scholarly works

Table I
CoP Dataset

| Name of the CoP | Members | Publications | Grants |
|---|---|---|---|
| Smart Cities | 10 | 320 | 323 |
| Cyber-Security | 28 | 404 | 380 |
| Neuroscience | 14 | 175 | 345 |

of a researcher or a faculty member. Among other things, a representation can include keywords, concepts, features, mathematical models, probabilistic model, or vector-space model retrieved from the scholarly data. The proposed work presents a few approaches for generating research footprints based on diverse scholarly data. In particular, we generate three footprints for each researcher, $r$, based on his/her publications (Def 1), submitted grant proposals (Def 2), and merged data comprised of all of his/her publications and proposals (Def 3).

$$pub\_footprint_r = \{pub_1, pub_2, ...., pub_n\}, \quad \text{(Def 1)}$$

where $pub_i$ is a tuple defined as $(title, Author, abstract, Keyword)$, where *Author* is a set of names, *Keyword* is a set of words, and $r \in Author$.

$$gr\_footprint_r = \{gr_1, gr_2, ...., gr_m\}, \quad \text{(Def 2)}$$

where $gr_i$ is a tuple that consists of $(title, abstract, principle\text{–}investigator, Co\text{–}investigator)$, where $Co\text{–}investigator$ is a set of names and $(r = principle\text{–}investigator \vee r \in Co\text{–}investigator)$.

$$merged\_footprint_r = (pub\_footprint_r, gr\_footprint_r) \quad \text{(Def 3)}$$

Described next are the details of different approaches for generating research footprints.

*1) Keyword-based Research Footprint:* Text analytical processes often rely on selecting a set of keywords to represent a document. Selection of keywords can vary based on the requirement of the analyses. While some approaches consider frequencies (Term Frequency Inverse Document Frequency), others use bag-of-words (BOW), or parts-of-speech tagging to extract keywords.

The proposed work utilizes Rapid Automatic Keyword Extraction (RAKE) [19] to automatically select a set of keywords from the title and abstract of each publication and grant. RAKE is an unsupervised and domain-independent method for extracting keywords. RAKE utilizes stop words and phrase delimiters to split document text into candidate keywords, which form sequences of content words that occur in textual information. Word co-occurrence is identified within these candidate words and finally, word scores are computed based on the degree and frequency. Extracted keywords, from publications and grants, are used to create keyword-based footprints for each faculty

*2) Concept-based Research Footprint:* The process of keyword extraction results in a set of keywords which may have different levels of abstraction. Given the short length of title and abstract, there is a possibility that the extracted keywords are less informative or potentially misleading. To address this challenge, we propose to generate footprints using concepts rather than keywords.

A concept is a high-level abstraction of related keywords, which may not be explicitly described in a document. There are a number of approaches available for deriving meaning from keywords including clustering [20], topic modeling [14], semantic similarity [21], concept identification [22], and document summarization [23]. The proposed work uses two methods for concept detection. The first method uses Wikipedia Miner [24] which provides simplified access to Wikipedia. It allows users to extract Wikipedia's rich semantic information. Wikipedia Miner provides web services that leverage Wikipedia's structure to support semantic comparison of terms and concepts and the cross-referencing of documents with relevant topics to detect Wikipedia topics mentioned in documents.

The proposed work also uses Maui [17] to detect concepts related to a set of keywords. Maui automatically determines main topics in documents by extracting keywords from text with or without the use of a reference to a controlled vocabulary. Maui utilizes training data consisting of textual information that has been tagged with appropriate topics and leverages semantic information extracted from Wikipedia to tag documents based on their content wherein each tag corresponds to a topic in a given document. Using the concepts identified by Wikipedia Miner and Maui, we generate six concept-based research footprints for each faculty.

*3) Title and Abstract-based Research Footprint:* The title and abstract of scholarly documents usually introduce the theme of the documents at a higher level of abstraction. Keywords or concepts extracted from a small set of text data with high levels of abstraction may fail to identify the exact area of specialization of the scholarly document. In the third and last approach for generating research footprints, we address this challenge by using the entire title and abstract of documents for generating footprints.

### C. CoP-Specific Model Training

In the study, we examine three CoPs: Smart Cities, Cyber-Security, and Neuroscience. These groups are selected from a set of 13 communities of practice at the university because each group has at least 10 members, and each CoP has the potential for interdisciplinary research. The CoPs have open membership, that is the researcher determines if s/he wants to join the CoP. The members of a CoP may or may not have expertise related to the CoP, which makes effective CoP alignment difficult in some scenarios. Each CoP also has a brief abstract and a set of keywords to describe the area of interest of the CoP.

CoP-specific model training focuses on generating a model for each CoP. The dataset used for training the CoP-specific models consist of one thousand publications extracted from ArXiv for each of the Smart Cities and Cyber-Security CoP's and 773 publications for the Neuroscience CoP. The publications are retrieved using query terms similar to the name of the CoP: smart cities, cyber security, and neuroscience. Each publication consists of title and abstract. Later, supervised machine learning methods

are applied on these publications to generate CoP-specific statistical models. The machine learning methods include three classifiers: Naive Bayes, Random Forest, and J48. These classifiers were selected to study the impact of data model (e.g., Naive-Bayes) and algorithmic model (e.g., C4.5, Random Forest) on generating CoP-specific model [25].

Naive Bayes classifiers depend on Bayes' theorem with the "naive" assumption that every feature is independent of any other feature that allows the probabilistic prediction of a class. Naive Bayes calculates a set of probabilities by counting the frequency and combinations of values in given dataset [26].

Random Forest classifiers use multiple learning algorithms to increase the predictive performance of classification and regression by formulating a multitude of decision trees during training, yielding the class that appears most often for classification or the mean prediction for regression [27]. J48 is an open source Java implementation of the C4.5 algorithm, which is an extension of the ID3 algorithm utilized to induce classification rules by generating decision trees for statistical classification [28].

The Waikato Environment for Knowledge Analysis (WEKA) machine learning suite is utilized for training and testing of the predictive models used for CoP classification [29]. Default WEKA settings are used for all three classifiers.

Along with different types of classifiers, this work also employs two variants of the classification method. The first method generates one model for the three CoPs using all the publications extracted from ArXiv labeled respectively (e.g., smart-cities, cyber-security, or neuroscience). Thus, given one footprint as input, this method provides three likelihood scores, one for each CoP. Each score indicates how likely the footprint is to belong to that CoP. The second variant consists of one model per CoP wherein the training dataset was labeled accordingly with publications belonging to that CoP (e.g., "smart-cities") and publications from the other two CoP's labeled as not being part of that CoP (e.g., "not smart-cities"). Thus, each footprint passes through three CoP-specific models. A 10-Fold Cross-validation using Naive Bayes, Random Forest, and J48 classifiers are performed on both variants. Table II shows the correctly classified instances for each of the classifiers for the three CoPs with both classifier variants.

Table II
ACCURACY OF THE CLASSIFIER VARIANTS

| | 1 model | 3 models (Average) |
|---|---|---|
| Naive Bayes | 90.2% | 91.9% |
| Random Forest | 94.9% | 95.86% |
| J48 | 94.3% | 96.16% |

### D. CoP Alignment

The faculty expertise profile consists of three research footprints: publications-only footprint, grants-only footprint, and merged footprint that uses both publications and grants. There are three variations of each type of footprint depending on how these footprints are generated: keyword-based, concept-based, and full-text-based, using the title and abstract of scholarly documents.

Within the CoP Alignment component, the CoP classifier(s) as described in Section IV-C, takes a faculty's footprints (pertaining to each method applied) as input and provides a classifier score for each of his/her scholarly document, using three different CoP models generated using different classifiers and two variants of these models. The classification process yields the probability of how likely a publication or grant belongs to a CoP. The classifier score can range from 0.0 to 1, zero indicating no likelihood and 1 indicating the maximum likelihood of a footprint to belong to a given CoP. The proposed work studies the impact of different thresholds on the classifier score for classifying scholarly data with a CoP. For example, a threshold value of 80% would indicate that scholarly documents that have a classifier score of 80% or more are considered for further analyses.

The Alignment component then aggregates the results of all scholarly data for each faculty member and assigns him/her to CoPs based on CoP coverage of the documents. CoP coverage for a given CoP and a faculty member is the percentage of documents of the faculty that is classified as part of the CoP. This work does not provide any weight on the CoP coverage. Thus, if a faculty member has one document that is classified with any given CoP, he/she is listed as a potential member of that CoP. The WEKA JAVA API, with default settings, is utilized to develop the CoP Matching component.

## V. EVALUATION

The goal of the work is to identify potential membership of researchers into CoPs based on his/her research expertise derived from different scholarly activities. The paper presents different approaches for generating research footprints using publication and proposal data. The footprints of each faculty, generated using different approaches, are used to identify potential membership of the faculty in the three CoPs. The following subsections present the results of the experiments in more details.

### A. Keyword-based Footprints for CoP Alignment

The first set of experiments uses keyword-based footprints for each faculty. Table III shows the classification accuracy of the faculty members of the three CoPs when keywords from their scholarly data are used to represent their research expertise. Note that, each CoP has a set of members (Section

Table III
COP MATCHING ACCURACY USING KEYWORD-BASED FOOTPRINTS AND ONE CLASSIFIER.

| | | 0% | | | 70% | | | 80% | | | 90% | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | NB* | J48 | RF** | NB | J48 | RF | NB | J48 | RF | NB | J48 | RF |
| Smart Cities | Pubs | 0.489 | 0.340 | 0.362 | 0.468 | 0.340 | 0.277 | 0.468 | 0.340 | 0.255 | 0.468 | 0.277 | 0.213 |
| | Grants | **0.723** | 0.383 | 0.362 | **0.723** | 0.362 | 0.404 | **0.745** | 0.362 | 0.255 | **0.766** | 0.362 | 0.213 |
| | Grants & Pubs | 0.681 | 0.404 | 0.383 | 0.660 | 0.404 | 0.447 | 0.660 | 0.404 | 0.298 | **0.702** | 0.340 | 0.213 |
| Cyber-Security | Pubs | **0.723** | 0.596 | 0.617 | **0.723** | 0.596 | 0.532 | 0.681 | 0.596 | 0.596 | **0.702** | 0.574 | 0.596 |
| | Grants | **0.809** | 0.532 | 0.617 | **0.766** | 0.532 | 0.532 | 0.404 | 0.447 | 0.553 | **0.766** | 0.532 | 0.574 |
| | Grants & Pubs | **0.702** | 0.489 | 0.574 | **0.723** | 0.489 | 0.511 | **0.723** | 0.489 | 0.574 | **0.702** | 0.511 | 0.574 |
| Neuroscience | Pubs | 0.532 | 0.489 | 0.553 | 0.489 | 0.489 | 0.426 | 0.511 | 0.489 | 0.383 | 0.511 | 0.511 | 0.298 |
| | Grants | **0.872** | **0.766** | **0.723** | **0.872** | **0.766** | **0.723** | **0.872** | **0.766** | 0.511 | **0.872** | **0.702** | 0.319 |
| | Grants & Pubs | **0.851** | **0.766** | **0.787** | **0.851** | **0.766** | **0.745** | **0.809** | 0.574 | 0.574 | **0.851** | **0.702** | 0.319 |

*NB = Naive Bayes **RF = Random Forest

Table IV
COP MATCHING ACCURACY USING KEYWORD-BASED FOOTPRINTS AND THREE CLASSIFIERS.

| | | 0% | | | 70% | | | 80% | | | 90% | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | NB* | J48 | RF** | NB | J48 | RF | NB | J48 | RF | NB | J48 | RF |
| Smart Cities | Pubs | 0.489 | 0.277 | 0.340 | 0.489 | 0.277 | 0.277 | 0.468 | 0.277 | 0.234 | 0.468 | 0.277 | 0.234 |
| | Grants | 0.617 | 0.340 | 0.447 | 0.681 | 0.340 | 0.340 | 0.681 | 0.340 | 0.298 | 0.681 | 0.340 | 0.213 |
| | Grants & Pubs | 0.638 | 0.362 | 0.447 | 0.617 | 0.362 | 0.404 | 0.638 | 0.362 | 0.319 | 0.638 | 0.362 | 0.234 |
| Cyber-Security | Pubs | 0.681 | 0.532 | 0.574 | 0.681 | 0.532 | 0.574 | 0.681 | 0.532 | 0.532 | 0.681 | 0.532 | 0.596 |
| | Grants | **0.809** | 0.553 | **0.702** | **0.809** | 0.553 | 0.596 | 0.426 | 0.447 | 0.532 | **0.809** | 0.553 | 0.574 |
| | Grants & Pubs | **0.702** | 0.511 | 0.617 | **0.766** | 0.511 | 0.596 | **0.787** | 0.511 | 0.574 | **0.723** | 0.511 | 0.574 |
| Neuroscience | Pubs | 0.553 | 0.553 | 0.574 | 0.574 | 0.553 | 0.468 | 0.511 | 0.553 | 0.468 | 0.532 | 0.553 | 0.319 |
| | Grants | **0.830** | **0.723** | **0.745** | **0.830** | **0.723** | 0.660 | **0.851** | **0.723** | 0.574 | **0.851** | **0.702** | 0.489 |
| | Grants & Pubs | **0.830** | **0.745** | **0.830** | **0.851** | **0.745** | **0.723** | **0.809** | 0.617 | 0.553 | **0.872** | **0.745** | 0.511 |

*NB = Naive Bayes **RF = Random Forest

III). The accuracy of the classifier on detecting the correct members for a given CoP, denoted as $cp$, is calculated as:

$$Accuracy_{cp} = \frac{\sum TP_{cp} + \sum TN_{cp}}{\sum_{j=1}^{n} members(cp_j)},$$

where $TP$ is the number of correctly identified members of $cp$ (i.e., True Positive), $TN$ is the number of correctly identified non-members of $cp$ (i.e., True Negative), $members$ is a function that returns the number of members in a CoP, and $n$ is the total number of CoPs.

Any accuracy greater than or equal to 70% is highlighted in the table. The proposed approach also applies a threshold on the classifier score to emphasize the publications and grants with a higher probability of belonging to any CoP. Four thresholds are used for the rest of the experiments: 0%, 70%, 80%, and 90%. Thus, when the threshold is set to 80%, it indicates that we only consider publications and grants that have a likelihood score equal to or more than 80% for belonging to any CoP.

Table III shows that, with keyword-based footprints, at 0% threshold, for the Smart Cities CoP, grants are the best predictor of CoP membership compared to publications. Using grants as the input and Naive Bayes as the classifier, the accuracy of the CoP matching (i.e., alignment) is 72.3%. This is the highest score compared to the other two classifiers (i.e., Random Forest and J48). For the Cyber-Security CoP, only Naive Bayes is able to reach more than 70% accuracy for the different approaches for generating the research footprints using grants and publication data. Similar to Smart Cities CoP, grants provided better accuracy for detecting membership to Cyber-Security CoP (80%) than publications. Both Random Forest and J48 performed poorly in assessing the membership.

We observe somewhat similar results for the Neuroscience CoP. Grants again performed best in this case with a Naive Bayes classifier (87.2%). The performance of J48 and Random Forest reached more than 70% for the Neuroscience group with grants, for both the thresholds of 0% and 70%. Beyond these thresholds, the accuracy of both these classifiers drops below 70% for both grant and publications. At 90% threshold, the accuracy of Naive Bayes stays somewhat similar for all three CoPs, with a consistently good performance for grants data.

Higher thresholds show a mixed impact on the accuracy. When higher thresholds are applied, the total number of publications considered is reduced. Therefore, if publications that score high are incorrectly classified, the overall accuracy will suffer due to the lower number of publications available for classification. Similarly, if publications are classified correctly, the accuracy will improve at higher thresholds.

Table IV shows the results of similar experiments with three classifiers, one for each CoP, instead of one classifier (Table III). The overall performance is lower than when one

Table V
COP MATCHING ACCURACY USING CONCEPTS FROM WIKIPEDIA MINER AND ONE CLASSIFIER.

| | | 0% | | | 70% | | | 80% | | | 90% | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | NB* | J48 | RF** | NB | J48 | RF | NB | J48 | RF | NB | J48 | RF |
| Smart Cities | Pubs | 0.298 | 0.213 | 0.213 | 0.511 | 0.213 | 0.213 | 0.468 | 0.213 | 0.255 | 0.468 | 0.213 | 0.213 |
| | Grants | 0.298 | 0.234 | 0.234 | 0.638 | 0.234 | 0.234 | 0.681 | 0.234 | 0.255 | 0.681 | 0.234 | 0.213 |
| | Grants & Pubs | 0.298 | 0.255 | 0.255 | 0.638 | 0.234 | 0.234 | 0.638 | 0.234 | 0.298 | 0.638 | 0.234 | 0.213 |
| Cyber-Security | Pubs | 0.660 | 0.596 | 0.596 | 0.660 | 0.596 | 0.553 | 0.681 | 0.596 | 0.532 | 0.681 | 0.596 | 0.596 |
| | Grants | 0.489 | 0.468 | 0.468 | 0.681 | 0.468 | 0.489 | 0.426 | 0.447 | 0.574 | **0.809** | 0.468 | 0.596 |
| | Grants & Pubs | 0.574 | 0.426 | 0.426 | 0.617 | 0.447 | 0.468 | **0.787** | 0.447 | 0.574 | **0.723** | 0.447 | 0.596 |
| Neuroscience | Pubs | 0.362 | 0.426 | 0.426 | 0.404 | 0.426 | 0.468 | 0.511 | 0.426 | 0.362 | 0.532 | 0.426 | 0.298 |
| | Grants | 0.489 | 0.660 | 0.660 | 0.383 | 0.660 | 0.638 | **0.851** | 0.660 | 0.404 | **0.851** | 0.660 | 0.298 |
| | Grants & Pubs | 0.426 | **0.702** | **0.702** | 0.489 | 0.681 | 0.660 | **0.830** | 0.638 | 0.553 | **0.872** | 0.681 | 0.298 |

*NB = Naive Bayes **RF = Random Forest

Table VI
COP MATCHING ACCURACY USING CONCEPTS FROM WIKIPEDIA MINER AND THREE CLASSIFIERS.

| | | 0% | | | 70% | | | 80% | | | 90% | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | NB* | J48 | RF** | NB | J48 | RF | NB | J48 | RF | NB | J48 | RF |
| Smart Cities | Pubs | 0.489 | 0.213 | 0.213 | 0.468 | 0.213 | 0.255 | 0.468 | 0.213 | 0.255 | 0.468 | 0.213 | 0.213 |
| | Grants | 0.574 | 0.213 | 0.234 | 0.660 | 0.213 | 0.277 | 0.596 | 0.213 | 0.213 | 0.574 | 0.213 | 0.213 |
| | Grants & Pubs | 0.574 | 0.213 | 0.234 | 0.681 | 0.213 | 0.319 | 0.617 | 0.213 | 0.255 | 0.617 | 0.213 | 0.213 |
| Cyber-Security | Pubs | 0.681 | 0.596 | 0.553 | **0.702** | 0.596 | 0.553 | **0.702** | 0.596 | 0.596 | **0.702** | 0.596 | 0.596 |
| | Grants | 0.660 | 0.447 | 0.447 | 0.681 | 0.447 | 0.638 | 0.511 | 0.426 | 0.574 | 0.638 | 0.447 | 0.596 |
| | Grants & Pubs | **0.702** | 0.426 | 0.426 | **0.723** | 0.426 | 0.574 | 0.660 | 0.426 | 0.574 | 0.617 | 0.426 | 0.596 |
| Neuroscience | Pubs | 0.489 | 0.426 | 0.468 | 0.447 | 0.426 | 0.383 | 0.468 | 0.426 | 0.298 | 0.447 | 0.426 | 0.298 |
| | Grants | 0.638 | 0.681 | 0.681 | 0.638 | 0.681 | 0.447 | 0.617 | 0.681 | 0.319 | 0.574 | 0.681 | 0.298 |
| | Grants & Pubs | 0.660 | **0.702** | **0.702** | 0.681 | **0.702** | 0.511 | 0.681 | 0.617 | 0.511 | 0.574 | **0.702** | 0.298 |

*NB = Naive Bayes **RF = Random Forest

classifier is used. The system is able to identify members of Neuroscience with the highest accuracy. Grants provide better results than publications. Naive Bayes performs better than the other two classifiers.

### *B. Concepts-based Footprints for CoP Alignment*

Table V presents the results of membership detection using one CoP classifier when concepts are used to generate the research footprints of the faculty members. The concepts are extracted from Wikipedia using the keywords appearing in the scholarly data of the faculty members. The results indicate, at 0% threshold, J48 and Random Forest are able to classify the members of the Neuroscience CoP with 70% accuracy when merged data of publications and grants are used to generate the research footprints. At 70% threshold, none of the classifiers are able to achieve 70% accuracy, although Naive Bayes achieves 68% accuracy for the Cyber-Security CoP when grants are used to generate the footprints.

The accuracy starts to improve at higher thresholds. At both 80% and 90% thresholds, Naive Bayes reaches more than 80% accuracy for the Cyber-Security and Neuroscience CoPs. As for the type of scholarly data, grants again are linked to higher accuracy. At 90% threshold, using Naive Bayes classifier and grants data, the system is able to achieve more than 80% accuracy for both the Cyber-Security and Neuroscience CoPs.

Table VI shows the result of similar experiments using three classifiers, one classifier for each CoP. The overall performance degrades compared to when one classifier is used. The performance of the type of classifier is also mixed in this case. Naive Bayes performs well for Cyber-Security with publications and merged data, with a 70% threshold. J48 shows 70% accuracy for Neuroscience at different thresholds, but only with the merged grants and publication data. Random forest achieves 70% accuracy once in the Neuroscience CoP with grants and publications at 0% threshold. Similar to the keyword-based footprint generation approach (Tables III and IV), three classifiers perform poorly compared to one classifier.

Similar experiments conducted using Maui as a concept identifier yielded somewhat similar results as shown in Table VII. Maui concepts, with one classifier, performs better for the Neuroscience CoP than the other two CoPs. Grants are a common element of any accuracy over 70%. J48 and Naive Bayes performs equally for identifying potential members of Neuroscience CoP at 90% threshold. However, only Naive Bayes performs well for the Cyber-Security CoP, specially with a higher threshold and with grants data.

Table VIII shows the accuracy of using concept-based footprints to identify members of CoPs when each CoP has one classifier. In this instance, three classifiers consistently achieve 70% or more accuracy score for the Neuroscience CoP with grants and merged publications and grants data. Naive Bayes and J48 show promising performance across

Table VII
COP MATCHING ACCURACY USING CONCEPTS FROM MAUI AND ONE CLASSIFIER.

| | | 0% | | | 70% | | | 80% | | | 90% | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | NB* | J48 | RF** | NB | J48 | RF | NB | J48 | RF | NB | J48 | RF |
| Smart Cities | Pubs | 0.255 | 0.213 | 0.213 | 0.489 | 0.213 | 0.213 | 0.468 | 0.213 | 0.255 | 0.468 | 0.213 | 0.213 |
| | Grants | 0.468 | 0.298 | 0.277 | 0.596 | 0.298 | 0.340 | 0.681 | 0.298 | 0.319 | 0.681 | 0.298 | 0.234 |
| | Grants & Pubs | 0.340 | 0.277 | 0.277 | 0.596 | 0.277 | 0.340 | 0.638 | 0.277 | 0.362 | 0.638 | 0.277 | 0.234 |
| Cyber-Security | Pubs | 0.596 | 0.574 | 0.574 | 0.638 | 0.574 | 0.553 | 0.681 | 0.574 | 0.553 | 0.681 | 0.574 | 0.596 |
| | Grants | 0.362 | 0.532 | 0.553 | **0.723** | 0.532 | 0.617 | 0.426 | 0.404 | 0.489 | **0.809** | 0.532 | 0.596 |
| | Grants & Pubs | 0.489 | 0.489 | 0.511 | 0.553 | 0.489 | 0.574 | **0.787** | 0.489 | 0.553 | **0.723** | 0.489 | 0.596 |
| Neuroscience | Pubs | 0.362 | 0.468 | 0.447 | 0.426 | 0.468 | 0.468 | 0.511 | 0.468 | 0.426 | 0.532 | 0.468 | 0.298 |
| | Grants | 0.511 | **0.702** | 0.681 | 0.319 | **0.702** | 0.638 | **0.851** | **0.702** | 0.617 | **0.851** | **0.702** | 0.489 |
| | Grants & Pubs | 0.426 | **0.723** | **0.702** | 0.511 | **0.723** | 0.681 | **0.851** | 0.617 | 0.532 | **0.872** | **0.723** | 0.489 |

*NB = Naive Bayes **RF = Random Forest

Table VIII
COP MATCHING ACCURACY USING CONCEPTS FROM MAUI AND THREE CLASSIFIERS.

| | | 0% | | | 70% | | | 80% | | | 90% | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | NB* | J48 | RF** | NB | J48 | RF | NB | J48 | RF | NB | J48 | RF |
| Smart Cities | Pubs | 0.511 | 0.213 | 0.213 | 0.468 | 0.213 | 0.255 | 0.489 | 0.213 | 0.234 | 0.511 | 0.213 | 0.213 |
| | Grants | 0.638 | 0.319 | 0.340 | **0.745** | 0.319 | 0.404 | 0.638 | 0.319 | 0.277 | 0.660 | 0.298 | 0.213 |
| | Grants & Pubs | 0.660 | 0.298 | 0.319 | **0.723** | 0.298 | 0.447 | 0.681 | 0.298 | 0.298 | **0.702** | 0.298 | 0.213 |
| Cyber-Security | Pubs | 0.681 | 0.596 | 0.532 | 0.660 | 0.596 | 0.638 | 0.660 | 0.596 | 0.596 | 0.660 | 0.596 | 0.596 |
| | Grants | 0.660 | 0.553 | 0.596 | 0.681 | 0.553 | 0.553 | 0.426 | 0.404 | 0.532 | 0.681 | 0.553 | 0.574 |
| | Grants & Pubs | 0.681 | 0.511 | 0.511 | **0.702** | 0.511 | 0.553 | 0.638 | 0.511 | 0.596 | 0.660 | 0.511 | 0.574 |
| Neuroscience | Pubs | 0.468 | 0.426 | 0.489 | 0.489 | 0.426 | 0.340 | 0.447 | 0.426 | 0.362 | 0.426 | 0.426 | 0.298 |
| | Grants | **0.787** | **0.702** | 0.660 | **0.830** | **0.702** | **0.702** | **0.745** | **0.702** | 0.532 | **0.723** | **0.702** | 0.489 |
| | Grants & Pubs | **0.723** | **0.702** | **0.702** | **0.766** | **0.702** | **0.702** | 0.681 | 0.617 | 0.489 | **0.702** | **0.702** | 0.489 |

*NB = Naive Bayes **RF = Random Forest

varying levels of thresholds, whereas Random Forest performs well at lower thresholds. The comparative performance of three classifiers with Maui-derived concepts are different than the earlier two cases (Tables IV and VI) where the system with three classifiers performs poorly than one classifier.

One of the major differences between one and three classifiers is that one classifier is able to detect Cyber-Security CoP members with more than 70% accuracy at higher levels of threshold, whereas three classifiers fail to do so. Conversely, one classifier fails to identify Smart cities CoP members, but three classifiers are able to detect membership at 70% threshold with a Naive Bayes classifier and grants data.

### C. Full-Text-based Footprints for CoP Alignment

The last set of experiments considers the full text of grants and publications for generating research footprints and using those footprints for CoP matching. Table IX shows the accuracy of using unprocessed text from the title and abstract of grants and publications to predict CoP membership using one classifier. These experiments, similar to the earlier ones, consider three research footprints for each researcher: publications-only footprint, grants-only footprint, and merged footprint. Similar to earlier approaches, Naive Bayes outperforms the other two classifiers. The best accuracy is achieved using grants data.

The combination of Naive Bayes with grants data yields the best accuracy for Smart Cities CoP, ranging from 70% to 74%. A similar trend is visible for the Cyber-Security CoP where both grants and merged data of grants and publications performs well with the Naive Bayes classifier at 70%, 80%, and 90% thresholds. The best performance is achieved by the Neuroscience CoP for all the classifiers at lower thresholds. However, at higher thresholds Random Forest fails to achieve high accuracy.

Table X shows the accuracy of CoP membership detection using three classifiers, one for each CoP. This variant performs poorly compared to one classifier for the Smart Cities CoP. The system only achieves 70% accuracy for Smart Cities with grants data and Naive Bayes classifier at 80% threshold, whereas one classifier achieved similar accuracy for all thresholds (Table IX). The result for Cyber-Security is better for this variant. The accuracy is higher and the system is able to achieve more than 70% for all thresholds unlike the one classifier variant that failed to do so at 0% threshold. The result of Neuroscience is somewhat similar to one classifier variant.

## VI. DISCUSSION AND FUTURE WORK

As demonstrated by the experimental results, among the various approaches for building research footprints, full text comprised of title and abstract of scholarly data performs best in identifying potential members of interdisciplinary

Table IX
COP MATCHING ACCURACY USING TITLE AND ABSTRACT AND ONE CLASSIFIER.

| | | 0% | | | 70% | | | 80% | | | 90% | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | NB* | J48 | RF** | NB | J48 | RF | NB | J48 | RF | NB | J48 | RF |
| Smart Cities | Pubs | 0.489 | 0.340 | 0.362 | 0.468 | 0.340 | 0.255 | 0.468 | 0.340 | 0.234 | 0.468 | 0.277 | 0.213 |
| | Grants | **0.702** | 0.383 | 0.362 | **0.702** | 0.362 | 0.340 | **0.702** | 0.362 | 0.234 | **0.745** | 0.362 | 0.213 |
| | Grants & Pubs | 0.660 | 0.383 | 0.362 | 0.660 | 0.383 | 0.383 | 0.681 | 0.383 | 0.255 | 0.681 | 0.319 | 0.213 |
| Cyber-Security | Pubs | 0.660 | 0.574 | 0.596 | 0.660 | 0.574 | 0.553 | 0.660 | 0.574 | 0.596 | 0.660 | 0.553 | 0.596 |
| | Grants | 0.745 | 0.532 | 0.638 | **0.766** | 0.532 | 0.574 | 0.426 | 0.447 | 0.553 | **0.745** | 0.532 | 0.553 |
| | Grants & Pubs | 0.681 | 0.489 | 0.596 | **0.702** | 0.489 | 0.574 | **0.702** | 0.489 | 0.553 | **0.702** | 0.511 | 0.553 |
| Neuroscience | Pubs | 0.553 | 0.511 | 0.574 | 0.532 | 0.511 | 0.468 | 0.532 | 0.511 | 0.404 | 0.532 | 0.532 | 0.298 |
| | Grants | **0.872** | **0.745** | **0.702** | **0.872** | **0.745** | 0.681 | **0.851** | **0.745** | 0.511 | **0.851** | **0.702** | 0.340 |
| | Grants & Pubs | **0.851** | **0.745** | **0.766** | **0.872** | **0.745** | **0.702** | **0.809** | 0.553 | 0.553 | **0.830** | **0.702** | 0.340 |

*NB = Naive Bayes **RF = Random Forest

Table X
COP MATCHING ACCURACY USING TITLE AND ABSTRACT AND THREE CLASSIFIERS.

| | | 0% | | | 70% | | | 80% | | | 90% | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | NB* | J48 | RF** | NB | J48 | RF | NB | J48 | RF | NB | J48 | RF |
| Smart Cities | Pubs | 0.489 | 0.255 | 0.319 | 0.489 | 0.255 | 0.277 | 0.468 | 0.255 | 0.234 | 0.468 | 0.255 | 0.213 |
| | Grants | 0.638 | 0.340 | 0.468 | 0.681 | 0.340 | 0.319 | **0.702** | 0.340 | 0.277 | 0.638 | 0.340 | 0.213 |
| | Grants & Pubs | 0.638 | 0.340 | 0.468 | 0.638 | 0.340 | 0.383 | 0.638 | 0.340 | 0.298 | 0.617 | 0.340 | 0.213 |
| Cyber-Security | Pubs | 0.681 | 0.511 | 0.553 | 0.681 | 0.511 | 0.596 | 0.660 | 0.511 | 0.532 | 0.681 | 0.511 | 0.596 |
| | Grants | **0.787** | 0.553 | 0.638 | **0.787** | 0.553 | 0.596 | 0.532 | 0.447 | 0.553 | **0.787** | 0.553 | 0.574 |
| | Grants & Pubs | **0.745** | 0.511 | 0.596 | **0.745** | 0.511 | 0.617 | **0.745** | 0.511 | 0.574 | **0.723** | 0.511 | 0.574 |
| Neuroscience | Pubs | 0.553 | 0.553 | 0.574 | 0.553 | 0.553 | 0.447 | 0.532 | 0.553 | 0.468 | 0.532 | 0.553 | 0.298 |
| | Grants | **0.830** | **0.723** | **0.766** | **0.851** | **0.723** | 0.638 | **0.851** | **0.723** | 0.553 | **0.851** | **0.702** | 0.489 |
| | Grants & Pubs | **0.830** | **0.745** | **0.830** | **0.872** | **0.745** | 0.681 | **0.830** | 0.617 | 0.511 | **0.851** | **0.745** | 0.489 |

*NB = Naive Bayes **RF = Random Forest

research groups. Keywords also performed equally well for most CoPs. Among the three different types of research footprints, concept-based footprints performs worst in detecting CoP membership. The poor performance is constant regardless of how the concepts were extracted. This indicates, when using small text (e.g., title and abstract), automatically extracted keywords or full text could provide better accuracy than using concepts that are linked to those keywords.

Another point to note is that publications alone are not a strong indicator of potential membership into a CoP. Compared to publications, grants are consistently better at correctly detecting CoP membership. The impact of classifier threshold is mixed. At times higher thresholds are linked to better accuracy, while other times the impact of a higher threshold is not visible. Of the three classifiers used, Naive Bayes achieves the best accuracy in almost all cases. J48 has better accuracy than Random Forest in most cases. Within the three CoPs, the publications and proposals associated with the Neuroscience CoP are a better predictor of IDR group membership.

In most cases, a higher accuracy is achieved when only one classifier is used. This is contrary to the performance observed during the model construction, when the best option is three independent classifiers, one for each CoP (Table II). The drop in performance (from 90% to around 70-80%) is probably due to the fact that the data used to align members with CoPs (e.g., publications, grants) is different than the data used to train the classifiers (e.g., ArXiv publications).

In the future, we plan to address the level of CoP membership of members, i.e., strength of research interest, in more details using fuzzy membership functions. We will study the impact of other scholarly data, such as keywords found in scientific profiles from Academia[2] or Research Gate[3], in research footprint generation. In addition, we plan to investigate statistical approaches for building research footprints. Other areas of research include incorporating temporal information in the footprints, using more CoPs for testing, addressing the low performance of publication data using data fusion techniques, and experimenting with different CoP model building approaches using a robust set of features.

## ACKNOWLEDGMENT

This work is supported in part by the National Science Foundation under Grant No. HRD-1242122. Any opinions, findings, and conclusions or recommendations expressed in this paper are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

[2]https://www.academia.edu/

[3]https://www.researchgate.net/

## REFERENCES


[1] N. R. Council, *Convergence: Facilitating Transdisciplinary Integration of Life Sciences, Physical Sciences, Engineering, and Beyond*. The National Academies Press, 2014. [Online]. Available: https://www.nap.edu/catalog/18722/convergence-facilitating-transdisciplinary-integration-of-life-sciences-physical-sciences-engineering {Accessed on: 6-20-2017}

[2] N. R. Council, Ed., *Enhancing the Effectiveness of Team Science*. The National Academies Press, 2015. [Online]. Available: https://www.nap.edu/catalog/19007/enhancing-the-effectiveness-of-team-science {Accessed on: 6-20-2017}

[3] D. Stokols, K. L. Hall, B. K. Taylor, and R. P. Moser, "The Science of Team Science," *American Journal of Preventive Medicine*, vol. 35, no. 2, pp. S77–S89. [Online]. Available: http://dx.doi.org/10.1016/j.amepre.2008.05.002, {Accessed on: 6-20-2017}

[4] D. B. Krafft, N. A. Cappadona, B. Caruso, J. Corson-Rikert, M. Devare, B. J. Lowe, and V. Collaboration, "VIVO: Enabling national networking of scientists," in *Proceedings of the Web Science Conference, 2010*, 2010. [Online]. Available: http://journal.webscience.org/316/, {Accessed on: 6-20-2017}

[5] A. L. Vogel, K. L. Hall, S. M. Fiore, J. T. Klein, L. M. Bennett, H. Gadlin, D. Stokols, L. C. Nebeling, S. Wuchty, K. Patrick, E. L. Spotts, C. Pohl, W. T. Riley, and H. J. Falk-Krzesinski, "The team science toolkit," *American Journal of Preventive Medicine*, vol. 45, no. 6, pp. 787 – 789, 2013.

[6] J. Tang, J. Zhang, L. Yao, J. Li, L. Zhang, and Z. Su, "Arnetminer: Extraction and mining of academic social networks," in *Proc. of the 14th ACM SIGKDD '08*, 2008, pp. 990–998.

[7] J. Diederich and T. Iofciu, "Finding communities of practice from user profiles based on folksonomies," in *In Proceedings of the 1st International Workshop on Building Technology Enhanced Learning solutions for Communities of Practice (TEL-CoPs06)*, vol. 213, 2006.

[8] T. Bansal, M. Das, and C. Bhattacharyya, "Content driven user profiling for comment-worthy recommendations of news and blog articles," in *Proceedings of the 9th ACM Conference on Recommender Systems, RecSys*, 2015, pp. 195–202.

[9] F. Osuna, B. Gurijala, P. Esparza, M. Akbar, and A. Q. Gates, "A feasibility study of an approach to extend research footprints using disparate sources," in *AAAI workshop of Scholarly Big Data: AI Perspectives, Challenges, and Ideas*, 2016.

[10] I. S. Ribeiro, R. L. Santos, M. A. Gonçalves, and A. H. Laender, "On tag recommendation for expertise profiling: A case study in the scientific domain," in *Proceedings of the Eighth ACM WSDM*, 2015, pp. 189–198.

[11] Y.-B. Kang, Y.-F. Li, and R. L. Coppel, "Capturing researcher expertise through mesh classification," in *Proceedings of the Eighth International K-CAP*, 2015, pp. 6:1–6:8.

[12] K. Ehrlich, C.-Y. Lin, and V. Griffiths-Fisher, "Searching for experts in the enterprise: Combining text and social network analysis," in *Proceedings of the 2007 International ACM GROUP*, 2007, pp. 117–126.

[13] C. Wang and D. M. Blei, "Collaborative topic modeling for recommending scientific articles," in *Proceedings of the 17th ACM SIGKDD*, 2011, pp. 448–456.

[14] D. M. Blei, "Probabilistic topic models," *Communications of the ACM*, vol. 55, no. 4, pp. 77–84, Apr. 2012.

[15] J. Tang, S. Wu, J. Sun, and H. Su, "Cross-domain collaboration recommendation," in *Proceedings of the 18th ACM SIGKDD*, 2012, pp. 1285–1293.

[16] T. Tomokiyo and M. Hurst, "A language model approach to keyphrase extraction," in *Proceedings of the ACL 2003 Workshop on Multiword Expressions: Analysis, Acquisition and Treatment - Volume 18*, 2003, pp. 33–40.

[17] O. Medelyan, E. Frank, and I. H. Witten, "Human-competitive tagging using automatic keyphrase extraction," in *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 3*, 2009, pp. 1318–1327.

[18] S. D. Gollapalli and C. Caragea, "Extracting keyphrases from research papers using citation networks," in *Proceedings of the Twenty-Eighth AAAI Conference on Artificial Intelligence*, 2014, pp. 1629–1635.

[19] S. Rose, D. Engel, N. Cramer, and W. Cowley, *Automatic Keyword Extraction from Individual Documents*. John Wiley and Sons, Ltd, 2010, pp. 1–20.

[20] A. K. Jain, M. N. Murty, and P. J. Flynn, "Data clustering: A review," *ACM Comput. Surv.*, vol. 31, no. 3, pp. 264–323, 1999.

[21] J. J. Jiang and D. W. Conrath, "Semantic similarity based on corpus statistics and lexical taxonomy," *CoRR*, vol. cmp-lg/9709008, 1997.

[22] G. H. Bower and T. R. Trabasso, *Concept identification*, 1964, pp. 32–94.

[23] J. Carbonell and J. Goldstein, "The use of MMR, diversity-based re-ranking for reordering documents and producing summaries," in *Proceedings of the 21st Annual International ACM SIGIR*, 1998, pp. 335–336.

[24] D. Milne and I. H. Witten, "An open-source toolkit for mining wikipedia," *Artif. Intell.*, vol. 194, pp. 222–239, January 2013.

[25] L. Breiman, "Statistical modeling: The two cultures," *Statistical Science*, vol. 16, no. 3, pp. 199–231, Aug. 2001.

[26] P. Langley, W. Iba, and, and K. Thompson, "An analysis of bayesian classifiers," in *Proceedings of the Tenth National Conference on Artificial Intelligence, AAAI*, 1992, pp. 223–228.

[27] L. Breiman, "Random forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, October 2001.

[28] J. R. Quinlan, *C4.5: Programs for Machine Learning*. Morgan Kaufmann Publishers Inc., 1993.

[29] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten, "The weka data mining software: An update," *SIGKDD Explor. Newsl.*, vol. 11, no. 1, pp. 10–18, November 2009.