2010 IEEE Fourth International Conference on Semantic Computing

A Taxonomy-based Model for Expertise Extrapolation

Delroy Cameron*, Boanerges Aleman-Meza†, I. Budak Arpinar§, Sheron L. Decker§ and Amit P. Sheth*
* Kno.e.sis Center, Wright State University, Dayton, OH 45435, USA
† Rice University, Houston, TX, 77005, USA
§ LSDIS Lab, University of Georgia, Athens, GA 30602-7404, USA

***Abstract*—While many *ExpertFinder* applications succeed in finding experts, their techniques are not always designed to capture the various levels at which expertise can be expressed. Indeed, expertise can be inferred from relationships between topics and subtopics in a taxonomy. The conventional wisdom is that expertise in subtopics is also indicative of expertise in higher level topics as well. The enrichment of *Expertise Profiles* for finding experts can therefore be facilitated by taking domain hierarchies into account. We present a novel semantics-based model for finding experts, expertise levels and collaboration levels in a peer review context, such as composing a Program Committee (PC) for a conference. The implicit coauthorship network encompassed by bibliographic data enables the possibility of discovering unknown experts within various degrees of separation in the coauthorship graph. Our results show an average of 85% recall in finding experts, when evaluated against three WWW Conference PCs and close to 80 additional comparable experts outside the immediate collaboration network of the PC Chairs.**

***Index Terms*—Expert Finder, Taxonomy, Bibliographic Data, Collaboration Networks, Semantic Association**

I. Introduction

Finding experts is typically a subjective and intuitive process, influenced largely by trust among individuals. People discover experts based on interactions with them and such knowledge propagates from person to person *prima facie*, through a human-centered information diffusion [1]. However, given inconsistencies in human perceptions and isolation within social networks [2], such a referral system may not always be practical. Thus, software systems primarily aimed at systematically determining who is an expert, have gained impetus in recent years. The goal of such Expert Finder systems is to quantitatively discover humans believed to have exceptional knowledge, cognitive and/or physical ability related to executing relatively complex tasks.

A key challenge for Expert Finder systems is obtaining the datasets from which expertise can be adequately derived. In the past a variety of sources have been used for finding experts in scientific research, including Citation linkage [3], Curricula Vitarum (CVs) [4], Intranet applications [5], [6], Version Control Systems (such as SVNs) [7], [8], patents and even research grants. However, the unavailability of such high quality datasets as SVNs, patents and intranet records, make them unsuitable for large scale tasks. Likewise, metadata from research grants[1] provide information at particularly coarse levels of granularity. Citation link analysis based on citation counts is vulnerable to the infamous *Pied Piper Effect* [9] in which a theory in a highly cited manuscript may later be disproved. Furthermore, naive citation link analysis may result in expertise overfitting, unless Citation Sentiment Detection is used to smooth the count perturbations. While citation sentiment detection has been addressed elsewhere [10], with regards to estimating expertise, the question remains, is a citation positive, negative or neutral [11].

With the emergence of digital libraries such as Science Direct and DBLP, metadata from scientific publications have become a viable alternative for expert finder research. The old adage that 'the publications of a researcher is representative of his/her expertise' [12] remains a fundamental notion. Furthermore, given the progression of Semantic Web standards and technologies, additional avenues for expertise analysis have become available. For example, a researcher with expertise in a search algorithm such as the "PageRank" algorithm is quite likely an expert in the general area of "Web Search," since PageRank is deemed a subtopic of Web Search. By identifying the associated topics for a publication, and modeling such topics in a taxonomy, a means of *inexact expertise matching* can be achieved.

Bibliographic data from scientific publications contain person-centric information indicative of collaboration relationships between researchers. Chen in [13] argues explicitly that a thorough understanding of bibliographic data enhances our understanding of the real-world entities in the underlying coauthorship network encapsulated by such data. Analysis of collaboration relationships in Expert Finder systems therefore has implications on collaboration network expansion. For example, a PC Chair may be more interested in discovering unknown experts, since she may already be familiar with many existing experts in the field. To uncover unknown experts, discovery of semantic associations [14] within the coauthorship graph is critical. Since many research communities exhibit the small world topology evidenced by SIGMOD's coauthorship graph [15], discovering experts outside the existing network could potentially circumvent collaboration stagnation.

There is therefore interest in identifying top researchers using bibliographic data [16], for which the semantics of a taxonomy bears significance. Furthermore, since experts unknown to PC Chairs may potentially be overlooked [17] and the human effort required to discover them may be

[1]NSF Grant Search - http://www.nsf.gov/awardsearch/

978-0-7695-4154-9 2010
U.S. Government Work Not Protected by U.S. Copyright
DOI 10.1109/ICSC.2010.27

IEEE computer society

substantial, an Expert Finder system is important. Even given referrals from trusted sources, the absence of a mechanism for quantifying expertise introduces uncertainty, making it difficult for PC Chairs to have confidence in unknown counterparts. Trust propagation may show *decadence* rather than *transitivity*.

Many challenges must be addressed when using publication data for finding experts. The construction of a taxonomy of topics from which expertise can be inferred is not trivial. Semi-automated approaches to taxonomy creation for Computer Science topics [18] have yielded topic hierarchies of only limited quality. And while more generic techniques for constructing Hierarchical Topic Models exist, such as Latent Dirichlet Allocation (LDA), a complete Computer Science topic hierarchy is not forthcoming. Even given topic hierarchies that link papers to various topics, such as the ACM classification hierarchy [19], widespread adoption by multiple publishers has not been realized. In spite of such challenges, the benefits of using publication data for finding experts have been argued and demonstrated with a small dataset in [20]. In this work, we take the additional step of using the semantics of a taxonomy as a key enabler for finding experts and expertise, and semantic associations for detecting collaboration relationships. The contributions of this paper are therefore:

- We address the problem of finding experts by applying semantic techniques under the scenario of finding relevant reviewers for consideration for membership in a Program Committee for conferences (or workshops, etc). The main benefit of semantics lies in measuring expertise at finer granularity, through inexact matching of expertise.
- We propose a model for finding relevant experts potentially unknown to PC Chair(s), involving discovery of Collaboration Levels among experts groups by analyzing semantic associations.
- We demonstrate the effectiveness of this approach by comparing existing experts listed in PCs of past conferences with recommended experts discovered using our techniques.

The rest of the paper is organized as follows. In Section II we present our expertise model, followed by the methodology for using this model to quantify expertise in Section III. In Section IV we describe the dataset used in the evaluation, followed by the evaluation in Section V. We present related work in Section VI and conclusion and future work in Section VII.

II. Expertise Model

In any expert finder application, a fundamental question is often "What is an expert?" Indeed, agreement on who or what an expert is, is a highly subjective matter which may even become controversial. We develop an expertise model for finding experts by first defining the notion of an author, then an expertise profile and finally the notion of expertise itself, to crystallize our conceptualization of an expert. Given the bibliographic data, the underlying coauthorship network and the publication corpus, we define an author as follows:

Definition 1. *Let $\mathcal{B}$ denote a bibliographic dataset $\mathcal{B}$=$\{b_1, b_2, \ldots, b_n\}$, where b_i denotes a bibitem, and let also $\mathcal{P}$ denote a publication corpus $\mathcal{P}$=$\{p_1, p_2, \ldots, p_n\}$, where p_i is a publication. Then there exists a 1-to-1 mapping between a bibitem b_i and a publication p_i such that the cardinality $|\mathcal{B}|$=$|\mathcal{P}|$ holds. Let also $\mathcal{T}$ denote the set of topics associated with $\mathcal{P}$, where $\mathcal{T}$=$\{t_1, t_2, \ldots, t_m\}$. Then there exists an $\mathcal{M}$-to-$\mathcal{N}$ mapping among topics and publications and there also exists a set of authors $\mathcal{A}$=$\{a_1, a_2, \ldots, a_j\}$ in $\mathcal{P}$, such that each author a_i is associated with a subset of bibtems $\mathcal{B}_i$ and publications $\mathcal{P}_i$, each linked to some subset of topics $\mathcal{T}_i$. Hence, an author a_i can be represented by the set:*

$$a_i = \{\mathcal{B}_i, \mathcal{P}_i, \mathcal{E}_i\} \tag{1}$$

where $\mathcal{E}_i$ refers to the expertise profile of the author a_i based on mappings between her publications $\mathcal{P}_i$ and their associated topics $\mathcal{T}_i$.

Given this definition we define an expertise profile as follows:

Definition 2. *The expertise profile of an author a_i, denoted $\mathcal{E}_i$ is a function of her publications $\mathcal{P}_i$ and their associated topics $\mathcal{T}_i$ along with the publication impact factor λ_d, of the proceedings in which each publication appears.*

$$\begin{aligned}\mathcal{E}_i(\mathcal{P}_i, \mathcal{T}_i) = \{&< p_1, \lambda_1, [t_{1,1}, t_{1,2}, \ldots, t_{1,x}] >,\\ &< p_2, \lambda_2, [t_{2,1}, t_{2,2}, \ldots, t_{2,y}] >, \ldots\\ &< p_d, \lambda_d, [t_{d,1}, t_{d,2}, \ldots, t_{d,z}] >\}\end{aligned} \tag{2}$$

where x in $t_{1,x}$ denotes the total number of topics for the first publication p_i and d in $t_{d,1}$ denotes the dth publication. It follows that $t_{d,1}$ denotes the first topic of publication p_d.

We argue as proponents in the ongoing debate on the reliability of impact factors [21], since our results make a case for their use in the above definition. Hence, we define expertise as follows:

Definition 3. *The expertise of an author a_i, denoted by $\varepsilon_i(t_k)$ or simply ε_i, is a normalized sum, which is a function of a seed topic t_k, such that given each topic $t_{d,x}$ for each publication p_d of the author a_i and its associated publication impact factor λ_i, then:*

$$\varepsilon_i(t_k) = \frac{\sum_{d=1}^{\mathcal{D}} (\vee_{x=1}^{X} p_d(t_{d,x}))\lambda_d}{\mathcal{Z}(\mathcal{A})} \tag{3}$$

where t_k denotes the seed topic and $\mathcal{Z}(\mathcal{A})$ is the normalization factor, equal to the maximum expertise across all authors given the seed topic t_k. Hence, $\mathcal{Z}(\mathcal{A}) = \max \varepsilon_i(t_k)$. The function $p_d(t_{d,x})$ is a boolean-valued function such that:

$$p_d(t_{d,x}) = \begin{cases} 1 & \text{if } t_{d,x} \in \tau^* \\ 0 & otherwise \end{cases} \tag{4}$$

where τ^* is the set containing the seed topic t_k and all its descendants in the taxonomy.

We illustrate expertise computation using this model with an example in Figure 1 in the next section.

III. Approach

The approach to finding experts involves three tasks; 1) creating *expertise profiles* 2) computing researcher *expertise* and 3) ranking experts according to their *expertise scores*. The first task can be viewed as a semantic annotation problem, in which papers must be assigned relevant labels. The second task requires quantification of expertise, for which a mechanism for assigning weights to relevant publications is needed. The final task of ranking experts requires sorting researchers by expertise and potentially collaboration levels for examination by PC Chairs. We discuss these in turn below.

A. Expertise Profiles

Several important assumptions are necessary when creating expertise profiles. The first assumption is that each topic t_k for which a publication can be associated will participate in the profile. For example, if a paper has three topics, then three topic-value pairs will appear in the expertise profile according to Equation (2). This is dissimilar from classification approaches, in which a collection of labels is presented for consideration and reliable filtering heuristics must necessarily be applied to determine which terms will be selected as class labels. Second, we assume that each topic is equally relevant to a publication, i.e. $w_{t_1}=w_{t_2}=,\ldots,=w_{t_x}$ where w_{t_i} is the weight of a topic linked to a publication. We take this approach because it is difficult to estimate the importance of keyword terms based on mere ordering, without machine learning or NLP approaches for full text classification. In this work, we develop our model using bibliographic data, limited only to paper abstracts, not full text. A third assumption is that every coauthor is implicitly and equally linked to the topics of a publication. We take as further justification for this claim, the fact that the overall aggregation of all publications within a research group clearly establishes the degree of researcher competence, based on number of publications alone. A Professor, for example, will likely have more publications on a certain topic than his students, and thus greater expertise. We acknowledge however, that in reality the distribution of knowledge in a publication has wide variance. There typically exists a primary author and possibly many advisors whose knowledge far outweigh other authors or some combination thereof. Since this information is not forthcoming from publications in the corpus we make the obvious assumption that equal expertise exists among coauthors, relying on the publication counts to smooth the distribution. Hence, for all authors on a paper, equal weights have been assigned, i.e. $w_{a_1}=w_{a_2}=,\ldots,=w_{a_r}$. Finally, we assume that the bibliography metadata used in this work are in fact correct. That is to say that we do not perform tasks for identifying variants of researchers names within the dataset as mentioned in [22].

Algorithm 1 outlines the steps taken in creating expertise profiles based on these assumptions. The input parameters are an author URI, a list of her publications $\mathcal{P}_i$, a papers-to-topics map and the taxonomy. Step 1 initializes an empty expertise profile, and Steps 2-4 initialize an empty topic list and publication impact factor for a given publication. If the proceedings

Algorithm 1 Create Expertise Profile

1: $\mathcal{E}_i \leftarrow \emptyset$ initialize empty expertise profile
2: **for all** publications p_d of an author a_i **do**
3: $\quad\tau_d \leftarrow \emptyset$ initialize empty list of topics for p_d
4: $\quad\lambda_d \leftarrow 0$ initialize publication impact factor to zero
5: $\quad$get 'publication impact' λ_d for p_d
6: $\quad$**if** $\lambda_d = null$ **then**
7: $\quad\quad\lambda_d \leftarrow$'default'
8: $\quad$**end if**
9: $\quad$get 'list-of-topics' $\mathcal{L}_d$ for p_d
10: $\quad$**for all** topics $t_{d,i}$ in $\mathcal{L}_d$ **do**
11: $\quad\quad$update τ_d with $t_{d,i}$
12: $\quad\quad$then recursively get each parent topic $h_{t_{d,i}}$ of $t_{d,i}$
13: $\quad\quad$**if** τ_d does not contain $h_{t_{d,i}}$ **then**
14: $\quad\quad\quad$update τ_d with $h_{t_{d,i}}$
15: $\quad\quad$**end if**
16: $\quad\quad$update $\mathcal{E}_i$ with $\{p_d, \lambda_d, \tau_d\}$
17: $\quad$**end for**
18: **end for**
19: **return** $\mathcal{E}_i$

for the article is absent from the dataset, an arbitrary default value of 0.10 is assigned in Step 7. By inspection, given the distribution of impact factors and particularly since most highly weighted and relevant proceedings to our evaluation in fact did appear in the dataset, this value applies only a small penalty, if not a small bias, to unknown proceedings. In Steps 9-16, all topics for which a paper is related, are obtained using the papers-topics-dataset and the taxonomy. Techniques for mapping papers to topics are covered in Section IV-C. Note that the topics in this list $\mathcal{L}_d$, may be subtopics in the taxonomy. Hence to expand an author's expertise profile, it is necessary to identify for each publication, all parent topics $h_{t_{d,i}}$ for a given topic $t_{d,i}$ and add both the topic and its ancestors to the topic list where appropriate. The final result yields an expertise profile in Step 19.

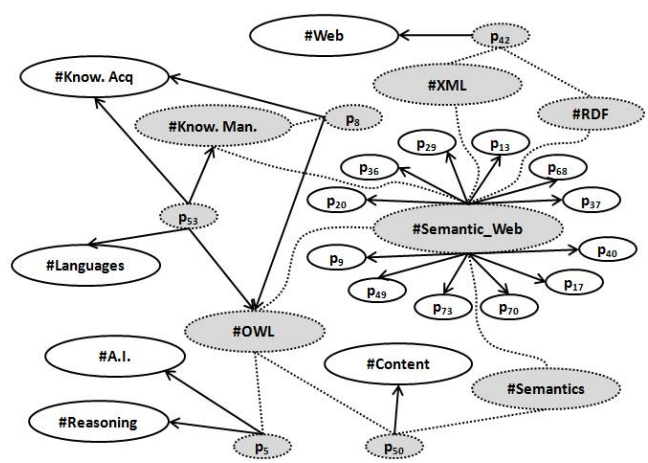

Figure 1. Snippet of an Expertise Profile taken from the dataset

The immediate benefit of this topic extrapolation is evident from Figure 1, which shows a snippet of an author's expertise profile taken from the dataset. Of the 81 total publications of the author, 12 can be mapped directly to the seed topic "Semantic Web," i.e. the nodes directly connected to the oval labeled #Semantic_Web in the graph. After generating the expanded expertise profile using Algorithm 1, an additional 5 publications (p_5, p_8, p_{42}, p_{50} and p_{53}) indicate expertise in Semantic Web, based on the topics OWL, Knowledge Management, Semantics, XML and RDF, each of which is related to Semantic Web in the taxonomy through some specialization relationship. It is intuitive then that the set τ^* from Equation (4) contains these and additional subtopics of the seed topic *'Semantic Web'*, i.e., τ^*={*Semantic Web OWL, Knowledge Management, Semantics, XML, RDF,* ... }. Manual inspection of the 5 publications confirmed that they were correctly identified as Semantic Web related, confirming that without extrapolation using the taxonomy, this information could have been overlooked and certainly would have been difficult to obtain otherwise. This result suggests that for this author, our technique has successfully realized a 30% improvement in recall assuming a total of 17 publications related to Semantic Web.

B. Computing Expertise

Computing expertise requires Equations (3) and (4) together. Considering again the expertise profile from Figure 1, the *aggregated expertise* of the researcher considering the 5 publications from which expertise must be inferred, can be computed as follows using Equation (3) that:

$$\begin{aligned}\varepsilon_i'' &= (p_5(OWL) \vee p_5(Reasoning) \vee p_5(A.I.))\lambda_{ecai} \\ &+ (p_8(Know.Acq) \vee p_8(Know.Man.) \vee p_8(OWL)\lambda_{ekaw} \\ &+ (p_{42}(XML) \vee p_{42}(RDF) \vee p_{42}(Web))\lambda_{www} \\ &+ (p_{50}(Semantics) \vee p_{50}(OWL) \vee p_{50}(Content))\lambda_{ewimt} \\ &+ (p_{53}(Languages) \vee p_{53}(Know.Acq) \vee p_{53}(OWL) \\ &\vee p_{53}(Know.Man.))\lambda_{ekaw}\end{aligned}$$

where $t_{5,1}$ denotes the first topic for publication p_5, such that $t_{5,1}$=OWL. For simplicity, since $p_5(t_{5,1}) \equiv p_5(OWL)$, we use the R.H.S. notation. Hence, for p_5 since OWL is a subtopic of Semantic Web (in τ^*), according to Equation (4), $p_5(OWL)$=1. Since A.I. and Reasoning are not, then $p_5(A.I.)$=$p_5(Reasoning)$=0. Given that λ_{ecai}=0.69, λ_{ekaw}=0.55, λ_{www}=1.54 and λ_{ewimt}=0.1 are the impact factors associated with each of the five publication proceedings, then from Equation (3):

$$\begin{aligned}\varepsilon_i'' &= (1 \vee 0 \vee 0)0.69 + (0 \vee 1 \vee 1)0.55 \\ &+ (1 \vee 1 \vee 0)1.54 + (1 \vee 1 \vee 0)0.1 + (0 \vee 0 \vee 1 \vee 1)0.55 \\ &= 3.43\end{aligned}$$

Given that the raw *aggregated expertise score* without the use of the taxonomy (i.e., using only the 12 publications directly related to Semantic Web) is ε_i'=10.4 (precomputed), then the overall aggregated expertise score of the author without normalization, is the sum of the aggregated scores for topic and subtopic matches, computed as ε_i^*=ε_i'+ε_i''. Hence, ε_i^*=10.4+3.43=13.83. The overall expertise of a_i with normalization, ε_i=$\frac{\varepsilon_i^*}{\mathcal{Z}(\mathcal{A})}$, where the precomputed maximum aggregated expertise score for the seed topic *Semantic Web* across all authors is, $\mathcal{Z}(\mathcal{A})$=32.23, is therefore ε_i=$\frac{13.83}{32.23}$=0.43. When compared with an expertise score excluding the taxonomy of only ε_i=$\frac{10.4}{32.23}$=0.32, it is evident that our methodology for computing expertise creates significant improvements in estimating expertise. We evaluate the recall in finding experts using this model in Section V.

C. Ranking Experts

Ranking experts based on our model involves sorting experts based on their overall expertise scores, which is fairly straightforward. In practice however, determining whether a researcher is an expert is a highly subjective matter. Ranking researchers could even become controversial. Such a scenario may arise due to various related disciplines comprised of very specific areas. For example, a researcher may be an expert in 'Database Systems' but not necessarily in 'Semantic Web.' This imposes limitations on making blanket statements about experts and expertise. Identifying experts requires a specific scope and well defined levels, to avoid misrepresentation. Our evaluation, uses a confined scenario to evaluate recall, owing to the absence of gold standards for ranking experts. While *h-index* for example, is a recognized benchmark for ranking Computer Science researchers, it is devoid of subject categories and therefore of less benefit than currently being sought.

D. Collaboration Networks

Our application for finding experts should provide the important functionality of alleviating the job of the PC Chair(s), who are themselves experts who in all likelihood are personally and/or professionally connected to other experts. The aim is therefore not only to find experts, but also to analyze collaboration networks relative to the PC Chair(s). The idea is to provide PC Chairs with information about who are the experts outside of their collaboration network, as a way to avoid possibly suggesting experts already known, i.e. "do not suggest to me experts whom I already know."

The notion of Semantic Associations [14] has been used to discover various paths that connect two entities in a populated ontology. This concept has applicability in a variety of areas, such as determining provenance and trust of data sources [23]. We implemented Semantic Association discovery to obtain multiple ways in which a given PC Chair is related to an expert. The goal is to automatically analyze each of these paths to determine the closest link between the two persons, called the Geodesic (γ).

1) Geodesic: We use the Geodesic (γ) as an enabler for PC Chairs' neighbourhood expansion. Table I describes the various levels of at which Geodesic relationships between two researchers can be expressed. γ_s is the strongest relationship, that of a coauthorship between authors. Slightly weaker relationships (γ_m) in which two authors may be connected by a common coauthor may also exist. Furthermore,

Table I
GEODESIC COLLABORATION LEVELS

Geodesic Level	Description w.r.t. PC Chair(s)	Degree/Sep.
STRONG(γ_s)	co-authors	One
MEDIUM(γ_m)	common coauthors	Two
WEAK(γ_w)	published in same proceedings	Unspecified
	coauthors w/ common coauthors	Two
	coauthor related to editor	Three
EXTR.WEAK(γ_e)	coauthors in same proceedings	Three
UNKNOWN(γ_u)	no relationship in dataset	Unknown

Table II
C-NET UNIT

Node (v)	Expertise (ε)	Collaboration Strength (c)
Super Node (v_m)	14.80	
Coauthor (v_1)	0.73	0.5
Coauthor (v_2)	0.73	0.5
Coauthor (v_3)	0.73	0.5
Coauthor (v_4)	1.81	1.0

even weaker relationships such as authors connected merely through publications in the same Proceedings (γ_w) or through coauthors with common coauthors (γ_e), also exist. Further still, a class of relationships is characterized as unknown (γ_u), since no data may exist in the dataset. Such relationships should certainly not be interpreted to mean that no relationship exists between the two researchers. Instead, personal and/or professional relationships may exist elsewhere, within other coauthorship networks distinct from that under consideration.

2) C-Net: In addition to Geodesic, we defined another type of collaboration relationship called 'C-Nets'. The C-Net paradigm borrows from that of "Relational ExpertiseNets" [3], in which analysis of incoming and outgoing citation links or ExpertiseNets, allude to the expertise of an author in various topics. In our model, C-Nets define the role of an author and his coauthors in a group, based on strong Geodesic connections amongst each other. For example, suppose that a Professor has three publications in a rather new topic and two of his students are coauthors in such publications. The goal is to use the measure of Collaboration Strength [12] within the group of Professor and students to distinguish which of these three persons might have higher 'authority' in that topic. In essence, a C-Net enables finding the "expert among closely connected experts."

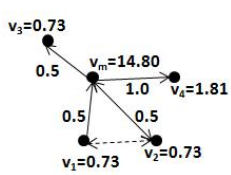

Figure 2. A C-Net Graph

Definition 4. *More formally, a C-Net is defined as a directed Graph $\mathcal{G}_k(\mathcal{V}_k, E_k)$, which is a subgraph of the coauthorship graph $\mathcal{G}(\mathcal{V}, E)$, in which there exists a super node v_m, that is connected with strong Geodesic (γ_s) to every other node within the subgraph such that v_m has maximum expertise (ε_m=ε_{max}) for a given topic t_x and the edge $e_j, e \in E_k$, between the super node v_m and any author, has a weight w_i equal to the collaboration strength c_i between the super node v_m and the author v_i, $v \in \mathcal{V}_k$, i.e. c_i=$w_j(v_m, v_i)$=$w_j(e_j)$.*

$$\textit{C-Net} = \mathcal{G}_k(\mathcal{V}_k, E_k) \tag{5}$$

Table II shows the C-Net for a subgraph from the dataset in which it is obvious that coauthor 4 (v_4) is more closely connected (c_4=1.0) to the super node v_m and has greater expertise (ε_4=1.81) on some topic dimension than her peers. Figure 2 shows a graph of the C-Net from Table II, in which directionality preserves order of coauthorship. That is, (v_1, v_m) indicates that the author v_1 has always been the first author and the super node v_m a coauthor and vice versa. In this graph, the order of coauthorship is maintained but does not affect the C-Net computations. In fact, we agree with the argument that the order in which a publication lists its authors, is generally not significant [24]. An additional advantage of reporting C-Nets is that a PC Chair may be interested to know who are the *rising researchers* (such as coauthor 4) within a field. It may be more enlightening to know that a student for example, is conducting cutting edge research in the area of 'Social Data for the Semantic Web' rather than 'Semantic Web' itself. The knowledge that the student is conducting research in the more general area of 'Semantic Web' is less informative.

IV. DATASET

The key dataset components for evaluating this expertise model are the 1) coauthorship network, 2) taxonomy of topics, 3) papers-to-topics dataset and 4) listing of publication impact factors. The coauthorship network is needed mainly for discovering semantic associations, which in turn is important for collaboration relationship detection and C-Net discovery. The network also contains the bibliographic data required for creating expertise profiles. The taxonomy allows expertise extrapolation, while the paper to topics dataset makes explicit, the links between taxonomy topics and publications. Impact factors aid in quantifying expertise. We discuss each of these datasets in this section.

A. Coauthorship Network

A coauthorship network is a graph in which vertices represent authors and edges connect coauthors. SwetoDBLP [25] is a coauthorship network that represents the DBLP bibliography in RDF, by semantically capturing its underlying connections through various relationships between articles, authors and proceedings. It contains 560,792 authors, related to over 900,000 articles, and was used in its entirety for constructing expertise profiles. However, in spite of promising research [26], finding semantic associations across large RDF graphs such as the SwetoDBLP[2] ontology is computationally expensive, for collaboration relationship detection we intuitively restricted our focus only to a small subset related to Web Search and Semantic Web. This subset contains only 67,366 authors, 25,973 journals and 51,202 conference proceedings.

[2]SWETO-DBLP - http://knoesis.wright.edu/library/ontologies/swetodblp/

Table III
SOURCE FOR TAXONOMY OF TOPICS

Source	# Topics
Conference Session Names	216
Conference Names	60
O'Comma Taxonomy	50
Paper Abstracts/Keywords	128
Total	320

B. Taxonomy of Topics

Our techniques for creating the taxonomy are manual but intuitive. For example, we assume that all publications appearing in a conference, such as the "International Conference on Semantic Computing (ICSC)" are relevant to the topic "Semantic Computing," which itself is a subtopic of Semantic Web. Semantic Computing would therefore be a topic linked explicitly to all papers appearing in ICSC and the relationship *Semantic_Web* $\longrightarrow$ *has_subtopic* $\longrightarrow$ *Semantic_Computing* would be maintained in the taxonomy. Such a subjective approach is not uncommon in the Semantic Web community. Often ontologies are created manually and refined through community agreement by domain experts. Since it is not our intent to create a comprehensive taxonomy of Computer science topics, we manually created this taxonomy related to Semantic Web and Web Search topics, consistent with our expertise. We initially populated this taxonomy with 216 topics obtained from session names from the related proceedings and later added an additional 60 topics obtained exclusively from conference names, both datasets obtained through focus crawling the DBLP web interface. An additional 128 topics were collected using keywords terms directly appearing in publications, then complemented using additional terms from paper abstracts, extracted using the Yahoo Term Extractor. Hyperlinks to external digital libraries such as ACM, IEEE and ScienceDirect, anchored by URLs in DBLP served as entry points to paper abstracts. Finally, an additional 50 topics (including some non Semantic Web related) were added from the O'CoMMA taxonomy [18] as a means of enrichment, refinement and possibly validation of our initial hierarchy. After various manual refinements the final taxonomy[3] consisting of 320 topics was produced. Table III summarizes topic contributions from the various data sources before curation. We note that many attempts that led to the development of taxonomies across a variety of areas have been undertaken [8], [18], however, to date no gold standard taxonomy exists for computer science topics. Hence it is difficult to independently evaluate the quality of the taxonomy we created with other known sources, beyond with what has been attempted here.

C. Papers to Topics Dataset

From an initial papers-to-topics dataset of 29,454 papers and 38,736 relationships linked through 128 topics obtained using the Yahoo Term extractor, we obtained an expanded papers-to-topics dataset containing 473,276 papers linked through 61,112 relationships, after integrating the topics obtained from the DBLP focus crawl for session names and the initial hierarchical relationships from the taxonomy. We obtained the complete dataset linking 476,299 papers through 676,569 relationships to 320 topics, after integrating the topics obtained from the second DBLP focus crawl that linked publications to conference names where appropriate. With this dataset, when applied to the SwetoDblp subset for collaboration discovery, there were 198,588 topic links among the 77,175 articles. Neglecting publications with single authors, each publication averages close to three (3) concrete topics in the taxonomy.

D. Publication Impact Factor

The final component of our dataset is Citeseer's publication impact statistics.[4] The advantage of Citeseer's publication impact listings is that it ranks over 1220 proceedings also cross-listed by DBLP, using DBLP conference URLs. The main disadvantage of Citeseer's impact factors however, is that this dataset is quickly becoming outdated. We take solace in the fact that our evaluation give credibility of these statistics based on recall.

V. EVALUATION

The core evaluation aims to determine the feasibility of the proposed expertise model for finding experts and collaboration relationships among them for aiding PC Chairs in selecting reviewers. A prototype Expert Finder application called SEMEF (SEMantic Expert Finder)[5] has been deployed and is available for browsing online.

A. Recall

To evaluate the recall of our model for finding experts we selected WWW Conference Tracks from 2005-2007, consistent with the latest version of the SwetoDBLP ontology. Notably, the applicability of our techniques is agnostic to the selected dataset.

First, author names from the WWW conference tracks were looked up in DBLP to obtain URIs for PC members, which enabled easy identification in the coauthorship graph, since SwetoDblp is DBLP in RDF. Expertise profiles for each PC member was created using Algorithm 1 described in Section III-A. The Call for Papers (CFP) for each track was used to obtain the initial seed topics. We manually obtained the initial seed topics for WWW2006 from the CFP as *Link Analysis, Crawling, Query Processing, Ranking, Indexing, Information Retrieval, Search* and obtained their subtopics from the taxonomy. Hence, using these seed topics ($\tau^*{}_{www}$) and also the precomputed expertise profiles for each PC member, expertise was computed according to Equations (3) and (4) from Section III-B. Ranking PC members within the PC-List was then straightforward based on their expertise scores.

The SEMEF-List (i.e. the unsupervised list of experts produced from our methods) was obtained by first iterating over all authors in the SwetoDblp subset and computing

[3]Taxonomy - http://knoesis1.wright.edu/expert_finder/swtopics/taxonomy

[4]Publication Impact - http://citeseer.ist.psu.edu/impact.html

[5]SEMEF Online Demo - http://knoesis1.wright.edu/expert_finder

Table IV
PROGRAM COMMITTEE LIST DISTRIBUTION IN SEMEF LIST

SEMEF %	Search Track # PC Members				Cumulative %
	2005	2006	2007	Avg.	
0-10	10	13	13	12	35
10-20	5	8	6	6	52
20-30	6	0	0	2	58
30-40	4	1	1	2	65
40-50	6	2	0	3	73
50-60	3	1	1	2	79
60-70	4	0	0	1	82
70-80	1	1	0	1	85
80-90	1	0	0	0	85
90-100	0	0	0	0	85
Total	40/48	26/29	21/25	29/34	
	83%	89%	84%	85%	

their expertise using the same seed topics and subtopics (i.e. $\tau^*{}_{www}$) then ranking experts based on their expertise scores. Disregarding researchers with infinitesimal expertise (less than 0.1) our SEMEF-List produced 540 researchers and more than 900 researchers when including them. Table IV shows the comparison between the ranked SEMEF-List of experts and the ranked PC-List of experts for the three conference tracks. The SEMEF-List on average produces 20 of 34 PC members in the first 30% (162 researchers) of SEMEF. Further, 35% of the PC-List (12 researchers) appeared in the top 10% (54 researchers) of SEMEF, while close to 60% of the PC are in the top 30% of SEMEF. Overall, SEMEF finds 85% of the experts in the PC-List to have greater than 0.0 expertise in the topics representative of the track based on our model and dataset. Hence, SEMEF achieves an 85% recall in finding these experts quantitatively. Additionally, on average, about 5 PC members for each track had zero expertise in the field. After manual investigation it was found that some papers were not linked to sensible session names, e.g. Session I, II etc. Additionally, broken and non existent hyperlinks in DBLP made it challenging to extract metadata from related online digital libraries that provide them. We speculated that some PC members could have been rising experts, such as PhD students but could not find any evidence of this without thorough exploration of PC member C-Nets, which was not undertaken.

B. Collaboration Network Expansion

The second aspect of our evaluation involved detecting semantic associations between researchers, given their expertise levels, presenting potentially unknown experts to PC Chairs for consideration. Table V shows collaboration relationships between PC Chairs-PC-List indicating that the majority of experts in the PC-List have a weak relationship to each Chair based on our dataset. We manually verified a small sample and found the STRONG and MEDIUM geodesic results to be consistent. Table VI shows that across the three tracks, on average there are 58 SEMEF researchers that have above average expertise and WEAK Geodesic with PC Chairs, when compared with the average expertise of the PC-List.

Table V
PC CHAIR - PC MEMBERS COLLABORATION RELATIONSHIPS

Geodesic	PC Chair - # PC members						
	2005		2006		2007		Above avg.
	C1	C2	C1	C2	C1	C2	
STRONG	2	0	3	0	3	0	0
MEDIUM	10	7	6	2	7	8	4
WEAK	31	17	15	20	11	14	10
EXTR. WEAK	1	2	1	2	0	0	0

Table VI
PC CHAIR - SEMEF LIST COLLABORATION RELATIONSHIPS

Geodesic	PC Chair - # SEMEF researchers						
	2005		2006		2007		Above avg.
	C1	C2	C1	C2	C1	C2	
STRONG	6	2	10	3	10	2	3
MEDIUM	106	53	88	55	88	76	16
WEAK	649	293	608	582	605	576	58
EXTR. WEAK	99	26	66	26	66	43	3

That is, an additional 58 experts connected through either having published in the same proceedings as the PC Chairs, or having common coauthor or some 3-hop relationship. In the real world, PC Chairs would be challenged to discover such experts through the proverbial *word-of-mouth* information diffusion. These findings strongly argue the benefit of our expertise model and the use of semantics-based techniques for expert finding. Given the fact that the benefits of C-Nets has already been demonstrated in Table II, the SEMEF paradigm potentially enables PC Chairs the flexibility of browsing the bibliographic dataset on the dimension of unknown experts, as well as other researchers in the immediate neighborhood of such unknown experts.

VI. RELATED WORK

In industrial settings, various approaches exist for finding experts. In [7], the concept of 'Expertise Atoms' is used to study expertise from version control data. The summation of code changes proves reliable in yielding expertise information. In this work we exploit publicly available datasets for finding experts, we have demonstrated proof of concept such that scalability is possible in theory given the appropriate resources. In [8] various algorithms for finding relevant reviewers are presented, based on coauthorship graphs and relative-rank particle-swarm propagation between coauthors. Edge weights propagate through the coauthorship network using stochastic analysis on outgoing edges, and state and energy levels of propagating nodes allow identification of the most qualified reviewers. Their approach finds qualified reviewers for the bidding phase of the peer review process and hence is a post activity once experts have been determined. The hierarchical ontological approach in [3], which classifies papers into expertise categories, bears some similarity to ours by using semantics on publication data. This similarity in approach is important in showing that there is value in data derived from

publications. Although this approach uses citation linkage and graph analysis to determine actual expertise values, we note the importance in using a taxonomy in relating publications to different topics is not prone to the Pied Piper Problem. Artnetminer [27] is similar to our work, but requires processing entire documents for obtaining expertise data. The difference is that we do not require full text precessing but only minimal information that will lead to the same conclusion. Additional work on author disambiguation which we do not address here is detailed in [28] based on using the underlying independent distribution in hidden topics in the publications of an author.

VII. Conclusion and Future Work

In this work, we presented a semantics-based expertise model and methodology for finding experts and expertise and we also demonstrated the value in using Semantic Associations for finding collaboration relationships in the context of the peer-review process. We examined collaboration networks between experts and PC Chairs, by exploring various Geodesic levels, as well as concepts called C-Nets for grouping experts. C-Nets enable the possibility of recognizing experts among experts as well as rising experts. In accomplishing these important tasks, we manually created a taxonomy comprising a subset of Computer Science topics useful for expanding expertise profiles through extrapolation of expertise. We evaluated our model by comparing experts found using SEMEF with PC members from three WWW conference tracks and found that SEMEF has 85% recall in the number of experts found. By inspection, we confirmed the correctness of Geodesic relationships for strong and weak associations. Additionally, we found a significant number of experts with weak relationships to PC Chairs as PC members, also with comparable expertise, hence producing candidates for PC consideration.

Future challenges include improving techniques mapping papers to topics through the use of machine learning. Similar approaches, such as LDA can be considered for creating a taxonomy or topic hierarchy across all topics in Computer Science. Also designing algorithms for finding semantic associations over large RDF graphs based on distributed cloud computing is necessary for finding semantic associations using larger datasets. Issues related to disambiguation are also necessary for improving the paper to topic mappings.

Acknowledgment

This research was supported by NSF grant No. IIS-0325464 awarded to the University of Georgia titled "Semdis: Discovering Complex Relationships in the Semantic Web" and is continued at Wright State University under Award No. 071441.

References

[1] M. Granovetter, "The strength of weak ties: A network theory revisited," vol. 1, 1983, pp. 201–233.

[2] M. Granoveter, "The strength of weak ties," in *American Journal of Sociology*, vol. 78, 1973, pp. 1360–1380.

[3] X. Song, B. L. Tseng, C.-Y. Lin, and M.-T. Sun, "Expertisenet: Relational and evolutionary expert modeling," in *User Modeling*, 2005, pp. 99–108.

[4] U. Bojars and J. G. Breslin, "Resumerdf: Expressing skill information on the semantic web," in *1st International Expert Finder Workshop*, 2007.

[5] P. Liu, J. Curson, and P. M. Dew, "Use of rdf for expertise matching within academia," *Knowl. Inf. Syst.*, vol. 8, no. 1, pp. 103–130, 2005.

[6] D. Mattox, M. T. Maybury, and D. Morey, "Enterprise expert and knowledge discovery," in *Proceedings of the 8th International Conference on Human-Computer Interaction*. Hillsdale, NJ, USA: L. Erlbaum Associates Inc., 1999, pp. 303–307.

[7] A. Mockus and J. D. Herbsleb, "Expertise browser: A quantitative approach to identifying expertise," in *In 2002 International Conference on Software Engineering*. ACM Press, 2002, pp. 503–512.

[8] M. A. Rodriguez and J. Bollen, "An algorithm to determine peer-reviewers," in *CIKM '08: Proceeding of the 17th ACM conference on Information and knowledge management*. New York, NY, USA: ACM, 2008, pp. 319–328.

[9] R. Kostoff, "The use and misuse of citation analysis in research evaluation," *Scientometrics*, vol. 43, pp. 27–43, 1998.

[10] S. Teufel, "echemistry: Science citations and sentiment," 2009.

[11] B. Pang and L. Lee, "Opinion mining and sentiment analysis," 2009.

[12] M. E. J. Newman, "Coauthorship networks and patterns of scientific collaboration," in *Proceedings of the National Academy of Sciences*, 2004, pp. 5200–5205.

[13] C. Chen, "Visualising semantic spaces and author co-citation networks in digital libraries," in *Information Processing and Management*, 1999, pp. 401–420.

[14] K. Anyanwu, A. Maduku, and A. P. Sheth, "Semrank: ranking complex relationship search results on the semantic web," in *In 14th International World Wide Web Conference*. ACM Press, 2005, pp. 117–127.

[15] M. A. Nascimento, J. Sander, and J. Pound, "Analysis of sigmod's co-authorship graph," *SIGMOD Rec.*, vol. 32, no. 3, pp. 8–10, 2003.

[16] M. Khalifa and K. Ning, "Demographic changes in is research productivity and impact," *Commun. ACM*, vol. 51, no. 4, pp. 89–94, 2008.

[17] D. Cameron, B. Aleman-meza, and I. B. Arpinar, "Collecting expertise of researchers for finding relevant experts in a peer-review setting," in *1st International ExpertFinder Workshop*, 2007.

[18] F. G. Acacia and F. Gandon, "Engineering an ontology for a multi-agents corporate memory system," in *Proc. International Symposium on the Management of Industrial and Corporate Knowledge 2001*, 2001, pp. 209–228.

[19] N. Coulter, "Acm's computing classification system reflects changing times," *Commun. ACM*, vol. 40, no. 12, pp. 111–112, 1997.

[20] S. Al Sudani, R. Alhulou, A. Napoli, and E. Nauer, "OntoBib: an Ontology-Based System for the Management of a bibliography," Research Report, 2006. [Online]. Available: http://hal.inria.fr/inria-00000972/en/

[21] P. O. Seglen, "Why the impact factor of journals should not be used for evaluating research," *BMJ*, vol. 314, no. 7079, pp. 497–, 1997. [Online]. Available: http://www.bmj.com

[22] D. Lee, J. Kang, P. Mitra, C. L. Giles, and B.-W. On, "Are your citations clean?" *COMMUNICATIONS OF THE ACM*, vol. 50, no. 12, pp. 33–38, 2007.

[23] L. Ding, P. Kolari, T. Finin, A. Joshi, Y. Peng, and Y. Yesha, "On homeland security and the semantic web: A provenance and trust aware inference framework," in *In AAAI Spring Symposium on AI Technologies for Homeland Security*. AAAI Press, 2005, pp. 21–23.

[24] B. Meyer, C. Choppy, J. Staunstrup, and J. van Leeuwen, "Viewpoint research evaluation for computer science," *Commun. ACM*, vol. 52, no. 4, pp. 31–34, 2009.

[25] B. Aleman-meza, F. Hakimpour, I. B. Arpinar, and A. P. Sheth, "A.p.: Swetodblp ontology of computer science publications," in *Journal of Web Semantics: Science, Services and Agents on the World Wide Web*, 2007, pp. 151–155.

[26] M. Janik and K. Kochut, "Brahms: A workbench rdf store and high performance memory system for semantic association discovery," in *In Fourth International Semantic Web Conference*. Springer, 2005, pp. 431–445.

[27] J. Tang, J. Zhang, D. Zhang, L. Yao, C. Zhu, and J. Li, "Arnetminer: An expertise oriented search system for web community," 2008.

[28] D. Mimno and A. Mccallum, "Expertise modeling for matching papers with reviewers," in *In Proceedings of the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2007, pp. 500–509.