Automated Expertise Retrieval: A Taxonomy-Based Survey and Open Issues

RODRIGO GONÇALVES and CARINA FRIEDRICH DORNELES,
Universidade Federal de Santa Catarina, Brazil

Understanding people's expertise is not a trivial task since it is time-consuming when manually executed. Automated approaches have become a topic of research in recent years in various scientific fields, such as information retrieval, databases, and machine learning. This article carries out a survey on automated expertise retrieval, i.e., finding data linked to a person that describes the person's expertise, which allows tasks such as profiling or finding people with a certain expertise. A faceted taxonomy is introduced that covers many of the existing approaches and classifies them on the basis of features chosen from studying the state-of-the-art. A list of open issues, with suggestions for future research topics, is introduced as well. It is hoped that our taxonomy and review of related works on expertise retrieval will be useful when analyzing different proposals and will allow a better understanding of existing work and a systematic classification of future work on the topic.

CCS Concepts: • **Information systems** → **Expert search**;

Additional Key Words and Phrases: Expertise retrieval, expert finding, expertise profile

ACM Reference format:
Rodrigo Gonçalves and Carina Friedrich Dorneles. 2019. Automated Expertise Retrieval: A Taxonomy-Based Survey and Open Issues. *ACM Comput. Surv.* 52, 5, Article 96 (September 2019), 30 pages.
https://doi.org/10.1145/3331000

1 INTRODUCTION

According to Balog et al. [11], *expertise* is a loosely defined concept that is not easily formalized or represented and usually referred to as *"tacit knowledge,"* i.e., the knowledge that people acquire through experiences in their lives that is stored in their minds. People can use this kind of knowledge to carry out tasks and solve problems, but it is difficult for them to express it in a detailed, formalized, and complete way that allows other people to know about their expertise. Finding ways to discover and automatically describe this type of knowledge is a valuable and challenging research topic.

One way to perceive tacit knowledge is to analyze the *expertise evidence* that is associated with a person. Expertise evidence includes any artifact from which information related to expertise can be extracted [11]. There are many sources from which these artifacts can be obtained: authored documents (articles, reports), electronic communications, and social networks, among others. The

Authors' addresses: R. Gonçalves (corresponding author), Universidade Federal de Santa Catarina, Campus Trindade, Florianópolis, SC, 88040-900, Brazil; email: rodrigo.g@ufsc.br; C. F. Dorneles, Universidade Federal de Santa Catarina, Florianópolis, Brazil; email: dorneles@inf.ufsc.br.

0360-0300/2019/09-ART96 $15.00
https://doi.org/10.1145/3331000

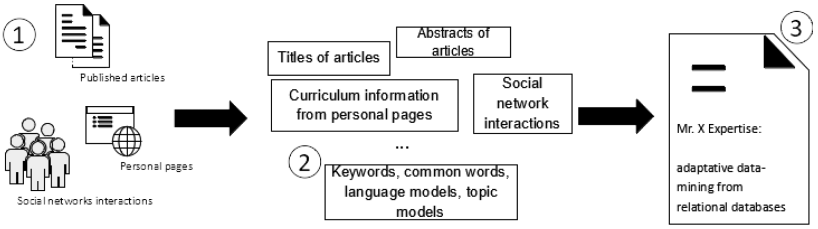
1
Published articles
Personal pages
Social networks interactions
Titles of articles
Abstracts of articles
Curriculum information from personal pages
Social network interactions
...
2
Keywords, common words, language models, topic models
3
Mr. X Expertise:
adaptative data-mining from relational databases

Fig. 1. The expertise retrieval process.

process of finding and extracting this kind of evidence and linking it to a certain expertise is called *expertise retrieval*, which is briefly summarized in Figure 1. In general terms, there are three basic stages:

(1) locating data sources for expertise evidence;
(2) extracting expertise evidence;
(3) making use of the evidence to formulate the person's expertise.

There are two basic applications for expertise retrieval: expert finding and expert profiling [11]. *Expert finding* focuses on a given list of one or more topics of interest and seeks to find experts related to these topics. *Expert profiling* is concerned with building expertise profiles, i.e., structured descriptions of people's expertise [9].

Automated approaches to retrieve expertise have become an interesting topic of research in recent years for many computer science communities: these include information retrieval (methods to extract expertise evidence and input from data, clustering, etc.) [19, 22, 27, 54, 56], databases (indexation and data structures) [30, 48, 117, 140], and machine learning (language models and topic models) [43, 66, 88, 92, 104, 107, 111, 136]. An automated expertise retrieval process basically follows the procedure described in Figure 1.

Following previous work [11, 89] and for a better understanding of the state-of-the-art with regard to automated expertise retrieval, we decided to conduct a survey on the topic, with the following goals:

- to characterize the existing work on four key components—data sources, data extraction, expertise representation and application, and provide a framework to enable their processes to be understood;
- to create a faceted taxonomy, based on the previous four components and to classify existing work (to the best of our knowledge, we are the first researchers to propose this kind of taxonomy);
- to analyze existing works, compare approaches, and discuss their benefits and drawbacks;
- to identify and discuss several open research issues with regard to automated expertise retrieval.

Compared to previous work [11, 89], our survey (i) introduces a faceted taxonomy (that can be extended and improved), serving as a guide for researchers to classify existing work and compare them, allowing, for example, selection of a subset of works to implement and compare for solving a given task; (ii) reviews existing work published up to 2017; and (iii) discusses several new open issues.

The remainder of this article is structured as follows. In Section 2, our proposed faceted taxonomy is introduced. Sections 3, 4, and 5 analyze the data source, data extraction, and expertise representation components of existing works, respectively. There is a discussion on the current applications for automated expertise retrieval in Section 6 and in Section 7 a comparison is made between the sampled work. This survey is concluded in Section 8, where there is an analysis of open issues and suggestions for further research.

2 TAXONOMY

In this section, a faceted taxonomy is created to classify existing work based on four chosen features. We start by introducing these features and later describe the taxonomy.

2.1 Overview

Expertise retrieval, as introduced earlier, follows three basic steps: (i) identify data sources from which expertise information can be retrieved, (ii) extract expertise evidence, and (iii) elaborate on a person's expertise. The data sources from which expertise evidence is extracted vary significantly and are introduced in detail in Section 3. The kind of information searched in these data sources includes title, keywords, abstracts, and text body from articles; text from documents stored in knowledge-management systems; messages in social networks; relations between people and what they have produced (citations, coauthoring network); and activities on the web (forums and question-and-answer sites, where aspects such as the body of answers and best-answer tags are considered), among others.

Once the data is localized, existing works apply many types of techniques to elaborate expertise evidences based on the data, i.e., some standard representation of the extracted data, the purpose of which is to provide information to allow executing expertise retrieval–related tasks. For example, language models [101] and topic models [17] are built over the text from abstracts and body of articles, documents, web pages, posts in SNS, and so on. Language models allow finding people whose documents are directly related to a given query (set of words). Topic models represent the probable expertise associated to a person through a summarized word-based representation of the person's production. Topic models can also be used to compare people, identifying those with similar expertise. Other works use the relations extracted from SNS and from researchers' production (such as coauthoring and citations) to locate those who stand out in a given topic (based on how many people refer to the person and how many publications the individual has). Section 5 introduces several examples of how existing works extract expertise information and Section 6 describes the tasks where this information is used.

After several expertise retrieval–related studies had been analyzed and their key components identified, we elected four key components to structure expertise retrieval processes:

- **Source:** This encompasses the data sources that can be used (and their linked features). We also examine how accessible they are (public unrestricted, public restricted, and private).
- **Extraction:** This represents the data extraction techniques that are designed in accordance with the inherent features of the data.
- **Representation:** This indicates the procedure used to produce knowledge about the expertise and employs extracted data; it can be regarded as the key stage in the process since most of the existing works introduce innovations. Many kinds of techniques were identified, such as language models, topic models, and graphs.
- **Application:** This is concerned with where the innovation is practiced; many studies have more than one application. Examples include ranking experts on the basis of a certain degree of expertise and profiling the expertise of professionals. Although introduced last, this component influences all others: the application enables one to define (a) which kind of data sources are needed, (b) the data to be extracted, and (c) what kind of knowledge is desired.

Table 1. Taxonomy Facets

Component	Facets
Data source	**Format:** unstructured, semi-structured, and structured **Accessibility:** public unrestricted, public restricted, and private **View:** plain text, communications, and dataset
Data extraction	**Expert composition:** simple and complex **Preprocessing:** none, word removal, and text transformation **Retrieval:** focused and complete
Expertise representation	**Method:** language model, term frequency, topic model, graph and custom **Temporal support:** none, time slices, and continuous **Semantic support:** none, ontology, lexical database, encyclopedia and knowledge database
Application	**Task:** expert finding, expert ranking, expert profiling, expert clustering, and expert recommendation

Following these components, we have created a taxonomy to classify the existing studies, which is outlined in the next section. We have designed a multifaceted taxonomy since many works cannot be classified by a single facet in every part of the taxonomy. The current studies are able to use more than one data source, apply different data-extraction methods, and have more than one application.

2.2 The Proposed Taxonomy

Our proposed faceted taxonomy is introduced in Table 1. This section briefly introduces its components; there will be a more detailed discussion in the next sections.

The taxonomy starts with the **data-source** component that contains facets describing its features: (i) *format* describes how their data is arranged, (ii) *accessibility* characterizes how accessible the data is, and (iii) *view* indicates how the data can be viewed.

The next component, **data extraction**, defines how existing works extract data from the data sources: (i) *expert composition* examines whether single or multiple semantic types of data are used to build a person's expertise; (ii) *preprocessing* is concerned with whether a given work executes some procedure with the extracted data to prepare it for future processing; and (iii) *retrieval* examines whether the existing work has gathered all of the available data on the data sources related to a person's expertise or just a subset, based on an initial query.

Expertise representation defines how the knowledge about expertise is formulated by means of the extracted data: (i) *method* defines the particular way the expertise is represented; (ii) *temporal support* is concerned with whether time is included in the expertise analysis—a person's expertise can vary during one's lifetime and, for example, outdated evidence of expertise may have to be rejected; and (iii) *semantic support* is concerned with whether the work uses semantic tools in the processes.

The last component in the proposed taxonomy, called **application**, contains facets related to the application, which requires the expertise retrieval process. Currently, it has a single facet, called *task*.

In the next sections, the components and their facets are analyzed in detail and some of the current work in the field is introduced. Owing to constraints with regard to space, the reader should refer to the original works for a more detailed description of their approaches.

3 DATA SOURCE

The **data source** component comprises three facets: *format*, *accessibility*, and *view*. Each facet is discussed in detail in this section.

3.1 Format

The *format* facet follows the well-known unstructured, semistructured, and structured classifications [1]. In this section, based on the surveyed studies, we exemplify expertise information sources for each classification.

3.1.1 Unstructured. *Unstructured* data includes data sources for which the semantics associated with the data is null or minimal. Account is taken here of any data source that does not impose semantics on its data. These data sources might include e-mail messages [2, 44, 134], scientific articles [69, 79, 116], wiki contents [110, 111], forums [95, 111], questionand-answer sites [29, 75], social network posts [79, 90], web pages [35, 79, 111] and knowledge management systems (KMSs) [135, 143].

When extracting data, unstructured data sources raise a number of challenges. If we find the name of someone at a prestigious university, this data can be subject to several interpretations. Perhaps this person works (or has worked) at this university or has coauthored a paper with someone else from the university, or it could also be that the person studied there, or it may even just be a citation of places where the individual would like to work or study.

Another common problem in unstructured data sources, but not limited to them, is, when faced with a document authored by several people, how do we know which person is responsible for which part? In the context of expertise retrieval, this problem is quite serious, since our objective is to find evidence of expertise so that a person can be profiled. In an article by Zhan et al. [146], for example, account is taken of the order of appearance of authors when assessing the significance of the expertise required for profiling authors. However, this may not be enough since the order of authorship does not always indicate to what extent a person is an expert in the subject of the article.

3.1.2 Semistructured. In *semistructured* data sources, there is support in their formats for introducing semantics to the data. Currently, the two main examples of semistructured data are XML and JSON documents.

In expertise retrieval, there are several cases where semistructured sources are used. For example, bibliographic databases, such as DBLP [78] and CiteSeer [109], publish their contents as XML files. DBLP and CiteSeer gather information from several sources and provide an integrated overview and search mechanism for this kind of data. Another example is AMiner [124], an academic search engine and mining system, which exports its data through an API (JSON) as well as semistructured text files.[1] Stack Overflow also publishes its data as XML files.[2] Semistructured data sources include publisher libraries such as ACM[3], IEEE[4] and Springer Link.[5] These allow some of their data to be exported in semistructured formats such as BibteX.

Some studies combine semistructured and unstructured data sources. Li et al. [79] and Fang et al. [43], for example, sought to retrieve expertise information through a person's scientific articles. The basic information about the articles—such as title, authorship, and keywords—was

[1] https://aminer.org/data.

[2] https://archive.org/details/stackexchange.

[3] http://dl.acm.org/.

[4] http://ieeexplore.ieee.org/Xplore/home.jsp.

[5] http://link.springer.com/.

retrieved from semistructured data sources. Unstructured data sources (the full text of the articles) were then queried so that further information could be extracted about the related expertise.

3.1.3 Structured. Structured data sources include those where the data has a well-defined, *stable*, and rigid format. The main example in this category is the relational database model.

To the best of our knowledge, there is no published work that explicitly uses structured data sources. This does not mean, however, that these data sources cannot be used for expertise retrieval work. Relational databases, from institutions such as universities or research centers, can include valuable information on a person's academic output and activities [93]. Retrieving and using this kind of data can greatly assist in designing an expertise profile.

3.2 Accessibility

The data sources used in current studies also vary in their level of accessibility. Accessibility refers to the amount of data that is published and how easy it is to access. Three levels were defined in our taxonomy: *public unrestricted*, *public restricted*, and *private*.

3.2.1 Public Unrestricted. These data sources provide an interface or a *dump* of their contents to external systems free of charge, either directly or through a previously created account. Through their interface, a computer agent can extract their data without restrictions. This is worth highlighting since there are cases where the data sources introduce limits to data extraction (such as *captchas* or rate limits).

Public unrestricted sources are mostly unstructured and semistructured, of which web pages are examples. In the context of expertise retrieval, there are personal home pages, project pages and institutional pages. Mailing list archives are another example of this level of accessibility. Data sources that are available in the Surface Web [14] are examples of public unrestricted data sources as well and include, for instance, public wikis such as Wikipedia.[6] Well-known examples of public unrestricted semistructured data sources include DBLP, CiteSeerX and the Stack Exchange network. DBLP publishes a *dump* of their data as a large XML file, which can be parsed and its information extracted. CiteSeerX provides an OAI (Open Archives Initiative Protocol for Metadata Harvesting) [76] interface. The Stack Exchange network publishes its data as XML files through the Internet Archive project.[7]

3.2.2 Public Restricted. These sources (common in the Deep Web), which provide access to their information, entail one or more of the following:

- licensing costs;
- providing restricted interfaces to extract data;
- limiting the available data.

Restrictions in interfaces to extract data include the need for human intervention (such as captchas) or limiting the volume of extracted data. Most publisher indices—such as ACM, IEEE, Springer, and Elsevier—are included in this category. ACM, IEEE, and Springer provide an interface to query and export the result (in BibTeX format) but are not computer friendly. Their policies, in some cases, explicitly state that computer-based harvesting of data is forbidden.[8] Other publishers, such as Elsevier, provide APIs to access their records but limit the results per query instance. These publishers restrict the available data as well—only subscribers can access the full text of the articles indexed for them, while nonsubscribers may access only metadata.

[6] http://www.wikipedia.org.
[7] https://archive.org/details/stackexchange.
[8] http://librarians.acm.org/policies.

The Lattes Platform[9] standardizes the curriculum format for Brazilian researchers. It is a central directory that allows researchers in Brazil to publish and update their profiles and include information such as published articles, books, participation in events, theses, and supervised learning. The curricula are accessible online through a search interface that requires providing a captcha since there is no public computer-agent, viewer-friendly interface.

Another example of services that are classified as public restricted data sources is social network systems (SNSs), such as Facebook, ResearchGate[10] and LinkedIn.[11] Facebook, for example, provides an API (called *Graphi API*) to extract limited data about people in its network. LinkedIn also provides an API to partners.[12] Twitter[13] provides an API, but it is limited, like Facebook.

Although public restricted data sources limit the amount of available information or the access rate, they are used in some studies [25, 88, 115, 116] since they provide valuable information. Some works combine public unrestricted and public restricted sources, using the former as a seed source and the latter as the source to be crawled. Fang et al. [43], for example, use data from AMiner together with Google Scholar[14] crawled abstracts.

3.2.3 Private. Private data sources do not provide public access. These sources are accessible only through internal networks in institutions/corporations. External users may not even know about their existence. Examples can be cited such as private KMSs [143], private SNSs [111] and micro-blogs [90]. Although not found in any of the studies reviewed, private wiki systems and HU (Human Resources) systems could equally be a private source of expert information. Intranet communications and e-mails [64, 151] are examples of possible data sources for expertise and are classified as private data sources as well.

3.3 View

The third facet of the data-source component is how the data can be viewed. Three types of data sources are proposed that can be classified as plain text, communications, and dataset.

Plain-text views are found in sources that are composed of unstructured documents and require data extraction methods to retrieve their data. The web pages available on the web or intranets (dynamic or static)—such as project pages, university pages, and personal pages—are examples of plain text [7, 13, 35, 111, 128]. Collections of documents in a corporation or institution are examples of plain-text data sources as well [45, 118, 143].

Communications views can be found in sources that represent message exchanging. This includes instant messaging [151], e-mails and mailing lists [6–8, 23, 39, 44, 117, 140, 152], web forums [7, 13, 111], and question-and-answer sites [15, 21, 29, 95]. SNSs [81, 90, 141, 142]—such as Facebook, ResearchGate, LinkedIn and Twitter—also provide communications views.

The concepts of *connections between people* and *exchange of information* are the main features that characterize a data source as having a communications view. Although in SNS there is clear information regarding personal relations (friends, acquaintances, and so on), from an expertise retrieval perspective, they can be seen as a communication system with additional information available. This information includes social interactions and relations.

Datasets involve providing data in a standardized and queryable format. This includes information systems, databases, and files in structured/semistructured format. The indexers of scientific

[9] http://lattes.cnpq.br/web/plataforma-lattes.
[10] https://www.researchgate.net/.
[11] http://www.linkedin.com.
[12] https://developer.linkedin.com/partner-programs.
[13] http://www.twitter.com.
[14] http://scholar.google.com/.

articles, such as ACM and IEEE, are examples of data sources that include a dataset view, since they provide either a web-based query interface or an API through which their data can be extracted [25, 43, 119, 148]. Wiki and KMSs provide dataset views as well—although they publish information as web pages, there is additional metadata that can be extracted [35, 111, 143] from their pages (such as author and abstract). Bibliographic databases (DBLP, CiteSeer[15], AMiner[16], ScholarMiner) are also examples of data sources with a dataset view [20, 22, 26, 30, 35, 43, 73, 79, 82, 100, 116, 119, 122, 126]. DBLP provides an XML dump[17] of its data. CiteSeer[18] and AMiner[19] have public APIs to extract data.

A data source can have more than one view. A wiki system, for example, introduces both a plain text and dataset view. The set of published pages is a plain text view while their associated metadata (internal to the wiki system) is a dataset view. The same happens with SNSs —if a system accesses public pages only from an SNS that originates from links in other pages, it will regard it as plain text. However, if it examines the relations between people through the SNS interface/API, it will see the same source from a communications view.

Another example of a multiple-view data source is a web forum or a questionand-answer site, composed of users and posts/replies. If there is access to the published pages resulting only from the thread to extract information, it is viewed as plain text. However, if the structure of questions and answers between users is taken into account, it can be seen as a communications data source.

4 DATA EXTRACTION

Automated expertise retrieval works vary on their method of extracting data from data sources. The three facets proposed to classify these methods are introduced in this section.

4.1 Expert Composition

Expert composition defines the extent to which more than one semantic type of data is used to represent expertise. By semantic type, we mean what kind of information the data represents. For example, textual words extracted from abstracts and from the body of scientific articles can be considered the same kind of information, i.e., keywords. On the other hand, although coauthor names and keywords are the same kind of data (textual words), they are not the same semantic types (keywords and names).

Two forms of expert composition are proposed: simple and complex. An item of *simple* data refers to a representation formed of one semantic type. For example, the expertise of a specialist can be represented as a set of words extracted from the specialist's papers, abstracts and titles or as the keywords associated with that individual's documents. *Complex* representations are formed of more than one semantic type. For example, if the keywords linked to an expert's documents are clustered by year, this is a complex representation since we have two semantic types of data (keyword and year).

Most of the studies that introduce expert composition based on terms extracted from linked documents are simple compositions [7, 10, 22, 30, 35, 47, 56, 62, 66, 79, 100, 102, 115, 128, 135]. Some [20, 33, 104] build a complex composition by linking each term to a moment in time, usually the publication year of the document.

Some studies use simple compositions based on the relations between experts (coauthoring, citations, and social relations) [70], while others introduce complex composition based on

[15] http://citeseerx.ist.psu.edu.
[16] http://aminer.org/.
[17] http://dblp.uni-trier.de/xml/.
[18] http://csxstatic.ist.psu.edu/about/data.
[19] http://doc.aminer.org/en/latest/.

coauthoring and temporal factors [55, 82]. Related entities are also used, such as the venues where an author's work has been published [73]. Combining both concepts extracted from the documents and coauthoring information [140] drawn on to represent an expert is another approach used in literature. Parada et al. [112] combine different features to represent the range of the researcher's interests.

4.2 Preprocessing

Some studies employ preprocessing techniques for the extracted data to improve the expertise retrieval. To classify existing work, three categories are proposed: (i) those that do not apply preprocessing techniques, classified as *none* in the proposed facet; (ii) those that remove unnecessary words from the text, designated as *word removal*; and (ii) those that transform the text, categorized as *text transformation*.

Stop words removal is the most common example of *word removal* technique. *Stop words* are previously known words that are very common (such as prepositions and articles) and do not contribute to a text's semantic value. Other examples of *word removal* include studies that remove words too rare or too common in the data being analyzed [34, 59, 66, 67, 104, 127]. Some remove punctuation and numbers as well [20, 144].

Stemming [120] reduces words to their root form and is the most used technique that performs *text transformation*. Johri et al. [62] normalize authors' names while analyzing their publications.

4.3 Retrieval

When extracting data, the existing studies retrieve all of the available data related to an expert (a *complete* retrieval) or just over a subset of the data (a *focused* retrieval), depending on the filters or conditions. Most of the studies that execute a complete retrieval do not provide results that are based on a given input/query; rather, they produce a result that can be browsed and analyzed by a user. This includes works that are designed to build researcher profiles [10, 25, 43, 47, 116, 126, 128] or conduct a coauthor network analysis [26, 30, 82, 140, 142] and research topic analysis [20, 62, 67, 73, 79, 132].

Focused retrieval works [37, 38, 44, 48, 85, 90, 98, 107, 111, 117, 121, 129] create a query with user input to define the subset of the data that needs to be extracted. They can be adapted to existing standard search engines with limited effort [7]. Most are deployed in expert finding.

5 EXPERTISE REPRESENTATION

Some existing studies employ techniques to build an expertise representation from data. Three facets are put forward here to display and classify them: *method*, *temporal support*, and *semantic support*.

5.1 Method

The methods for building expertise representation can be divided into five categories: term frequency, language model, topic model, graph, and custom. Although topic models and language models use term frequency as well since they introduce distinct features and possibilities compared with traditional term frequency methods, they were classified separately. In this section, each category is described in detail together with a brief explanation of their fundamental principles while the reader is referred to several studies.

5.1.1 Term Frequency. *Term frequency* uses the frequency of terms in a document to define or retrieve the expertise related to a person [10, 19, 22, 30, 35, 42, 56, 74, 135]. Its basic aim is to consider how many times a term appears in a given document so that its relevance can be assessed.

Term frequencies can be used to build vectors [135] that represent an author [19, 22, 30, 47, 54, 56] or a document [27, 90, 130]. Each term is a dimension in the vector. When building an author representation, for example, the contents of all related documents can be joined together and viewed as a single document [22] to calculate the frequencies and build the vector. Similarity metrics between vectors, such as cosine distance, are used to compare an expert with a given query (for expert retrieval) or another expert (for clustering [19] or collaboration recommendation [30]).

5.1.2 Language Model. According to Manning et al. [101], a language model is a "*function that puts a probability measure over strings drawn from some vocabulary.*" It is formed on the basis of an existing text by analyzing how frequently certain terms (a word or group of words) appear. Each term is assigned a relative frequency that is used to build the probability distribution model. Once the model is built, it can be used with a new text to calculate the probability that it formed a part of the data used to build the model. A language model built over the documents produced by an author, for example, indicates the degree of probability that a given text was written by that author [7, 9, 10, 43, 99, 107, 121, 142].

5.1.3 Topic Models. Topic Models [17] are probability distributions with regard to the topics of a given document. The assigned probability shows how probable some related material explores/contains information about the topic. One way to determine which topics should be related to a document is thorough its metadata, for example, a list of related categories. When seeking the topics of a given author, the categories related to each document the author has produced might be collected.

In expertise retrieval, topic model approaches can be adopted to identify topics by analyzing the contents of documents: for example, the abstracts of articles. Following classical topic modeling [17], topics are defined as clusters of related words obtained through statistical analysis. Each topic is represented by a set of words that can be defined through sampling methods such as those of Gibbs [49].

Topic models do not limit themselves to modeling topics arising with documents. Several studies introduce *hidden variables*, i.e., probability distributions with regard to the features in the domain. A hidden variable can represent, for example, (i) the probability of an author writing about a given topic, based on the topics of the author's related documents; and (ii) how likely it is that a given topic will be featured in a conference, based on articles from previous editions. The possibility of creating new *hidden variables* based on existing data (such as document topic and document-conference relations) allows a wide range of topic model methods to be devised. As a result, it was found to be the most common expertise evidence extraction technique in the studies that were reviewed.

There are several approaches to solving the problem of topic modeling. In the context of expertise retrieval, a well-known approach is latent Dirichlet allocation (LDA), introduced by Blei et al. [18]. This designs a generative probabilistic model for data collections through a three-level hierarchical Bayesian model, applicable to text corpora. The author-topic model, outlined by Rosen-Zvi et al. [122], also models the topic distribution for the authors concerned, i.e., by determining which topics are shared by each author. Other studies [24, 25, 41, 57–59, 73, 88, 103, 111, 137] extend the author-topic model by including additional features such as venues [28, 144], document citations [65, 127], preexisting supervised document subject classification [103], or cooperation information between authors to improve the efficiency of topic discovery [46]. Xie et al. [138] introduced a topic model that covers social interactions and relationships in social networks when building an expert's topic model and ranking the individual's expertise on a desired topic.

5.1.4 Graph. Graph techniques include work where a graph representation of expertise data is generated. The graph can be designed with the aid of the original data [37] or by using transformed data that are generated through another technique (topic and language models [25, 88, 142]). Once the graph has been generated, specialized methods are employed to extract the required information. *Page-rank -like* methods such as Random Walk [48, 117] are used for collaboration recommendation and expertise retrieval (i.e., expert ranking). Other studies [26] use the structure of the graph to calculate similarities between entities (such as authors) and recommend collaborations.

Some works [88, 113] use weighted graphs to represent, for example, coauthorship networks [113]. Other examples of graph-based techniques include (a) author profiling based on coauthors [53], (b) finding the most probable author for a given topic [88, 117] and the closest people to a given person [25, 142], or (c) clustering people in terms of their expertise [56, 73]. Kong et al. [71] and Xie et al. [138] combine topic model comparison between authors with a random-walk technique within a collaboration graph to suggest collaborations. De La Robertie et al. [36] present the RAC model, which uses previous information (conference authority) to help identify experts on a given topic by applying a label propagation algorithm. The *knowledge graph* has been proposed as well [96] as a tool for finding users to answer questions in Community Question Answer (CQA) sites.

5.1.5 Custom. This category includes studies that use alternative techniques. For example, Punnarut et al. [114] created a researcher profile based on an ontology formed through extracting terms from documents and matching them with a previously defined list of skills. Latif et al. [77] uses Linked open data (LOD) [16], which is available on the web, to build researcher profiles. LOD is a method based on standard web technologies such as HTTP, RDF, and URIs and is employed to publish data in a way that can be automatically interpreted and handled by computers.

Fang et al. [44] introduced a discriminative model that integrates documentary evidence of expertise and document-candidate associations in a learning framework for expert searching and ranking. Macdonald et al. [98] proposed using the voting model to rank experts from an expert search result. Ban and Liu [12] sought to combine graph techniques (using a citation network) and introduced a customized *** (Vector Space Model (VSM) which includes the location where the terms appear in an article so that they can be weighted) as a way to find experts.

5.2 Temporal Support

A person's expertise is not immutable, i.e., it changes over time owing to factors such as changes in the subject of interest or just a lack of continuity in previous interests. This means that the evidence a person has expertise on a given topic preferably should be viewed in the context of time.

Suppose, for instance, that there is a need for an expert in GIS databases. Two researchers are selected on the basis of their output, *ResearcherA* and *ResearcherB*. ResearcherA published many works 10 years ago but in recent years has focused his research on *distributed transactions*. ResearcherB started to publish papers on GIS databases only three years ago but has maintained a constant output. How can one choose between them? The temporal aspect of the expertise evidence may make it easier to make a decision on the basis of what activity is required from the expert. For lecturing undergraduate students about GIS databases, either ResearcherA or ResearcherB could be invited. However, with regard to integrating a new research project, ResearcherB should be preferred, since he will probably be more interested in this than ResearcherA, who has changed his research field in recent years.

Naturally, it is not a trivial issue to define the effects of time on the expertise evidence. Many studies [20, 25, 30, 33, 43, 50, 57, 61, 68, 85, 104, 107, 116, 137, 139, 145] include time as an important feature in their analysis as well as the extraction of expertise evidence. Three possible ways regarding *if and how* they incorporate time in their approaches can be distinguished: (i) *none*—time is not taken into account in their analysis, (ii) *time slices*, and (iii) *continuous*.

5.2.1 *Time Slices.*

In these approaches, the evolving pattern of expertise is analyzed in *slices* of time, such as a year, where each slice can be influenced by evidence from previous slices. For example, Chaiwanarom and Lursinsap [25] analyzed the evolving expertise of a researcher using that person's topics of interest over a period of time by sliding a window of a fixed number of years. This process produces a function that estimates the probable research interests in the future. Fang et al. [43] and Rybak et al. [116] analyzed topic evolution per year for a given author through probabilistic functions over time (where a given annual probability depends on previous years). Kong et al. [71] build per-year topic models (LDA) for authors' output by analyzing their dynamic research interests over a period of years.

Neshati et al. [107] examined the question of research longevity (in number of years) when estimating the strength of the relation between an author and published paper topics when the individual is a coauthor. Li et al. [85] used a time-partitioned random walk in a graph to analyze evolving expertise in a social network. Bolelli et al. [20] included the time factor in their topic model (S-ATM; segmented author-topic model) when they analyzed the topic evolution per time unit (year), in which previous years have a decaying influence on the current year. Daud et al. [33] proposed temporal-author-topic (TAT), which introduces a similar idea.

Jin et al. [60] analyzed the number of publications per topic and year to discover changes and tendencies in the expert's interests. Neshati et al. [106] analyzed the evolving topic model of experts based on question-and-answer sites. They introduced four features that affect topic transitions: (i) topic similarity—users usually change between similar topics; (ii) emerging topics—users tend to prefer emerging topics; (iii) user behavior—how common it is for a user to explore and change topics of interest; and (iv) topic transition—determining which topic changes are most common.

5.2.2 *Continuous.*

Some schemes do not require predefined time slices when the expertise evolution is being analyzed. Jameel and Lam [57], for example, designed a topic model based on n-grams, where each *topical phrase* has a timestamp associated with it and the expertise evolution is incorporated into the topic model itself. Naveed at al. [104] included absolute timestamps in their topic model (ATTention). Kawamae [68] also included timestamps in his theme chronicle model, and defined the concepts of stable and dynamic topics.

He et al. [50] analyzed topic evolution through citations between papers by taking account of time in their Inheritance Topic Model (ITM), which models documents as two parts that are generated independently: an *inherited* and *autonomous* part. The former is the outcome of previous work (based on the citations found). Wang et al. [137] included time as well when analyzing topic evolution in their citation-LDA topic model. Jo et al. [61] analyzed a collection of documents in chronological order and, in this way, established topic evolution. Estimating the future expertise of users in CQAs sites, including the transition probability to a new topic, has been researched as well [106].

Zehnalova et al. [145] devised a *forgetting function* to analyze topic evolution of an author over a period of time. Cohen et al. [30] examined the time that had elapsed while analyzing collaborations between authors in a coauthoring networking. Xu et al. [139] introduced the author-topic over time (AToT) model, a topic model that includes a timestamp associated with the topics used to design an author's interest model and its changes over time. Xie et al. [138] investigated the timestamps associated with microblogs from users as a sign that there were more interesting experts for a given user. This study was based on his own microblogs, timestamps, and Internet usage.

5.3 Semantic Support

The last facet in the expertise representation component of the taxonomy classifies existing works in terms of what kind of semantic support they use: none, ontology, lexical database and knowledge base. Most of the current studies do not use semantic support.

Among those that rely on ontologies, some require an ontology that has been prepared in advance [94, 116, 128] while others construct one during their processes [45, 63, 114, 140]. Those that rely on lexical databases use them to build ontologies that are based on word relations [63, 140] or to overcome problems regarding the usage of terms in documents (such as synonyms, hypernyms, and hyponyms) by finding equivalent words [19]. Two examples of knowledge bases used by researchers are DBpedia [110] and Wikipedia [27, 35].

This study classifies works that rely on Wikipedia in the *knowledge base* category since they use it as a support for their processes even though Wikipedia does not provide the semantic structure expected from a traditional knowledge database (as in DBpedia). For example, Davoodi et al. [35] build a vectorial representation of Wikipedia articles (based on term frequency) using the vectors to identify semantic topics in documents by comparing their vectorial representation.

6 APPLICATION

The *Application* that requires automated expertise retrieval can perform several kinds of *tasks*. In related work outlined here, we have identified five basic tasks: expert finding, expert ranking, expert profiling, expert clustering, and expert recommendation. In this section, each task is discussed together with information about the particular features of related work. In each application, we selected representative work to introduce a more detailed discussion.

6.1 Expert Finding

An *expert finding* procedure involves looking for experts through a search query. The query parameters vary with each proposal but most expect to find expertise topics as input. There are two basic expert finding approaches in the literature: (i) compiling a specialist index based on expertise-related information [27, 38, 81, 88, 107, 113, 125, 127, 129–131, 136] and (ii) using traditional indices (such as an inverted index) to locate documents related to a given expertise and employ expert finding methods on the basis of the results [44, 90, 92, 111]. Although most approaches look for a single expert, there are studies in the literature that focus on finding groups of experts as well [86, 105].

Current studies in the field adopt several approaches to finding experts. Many use document-centric methods, such as (i) using the SVM to represent and search for experts, given keywords of interest [27, 81]; (ii) constructing language and topic models based on a person's associated documents and a given input as a set of terms or topics for finding those experts whose models can best generate the query [27, 86, 88, 92, 102, 111, 125, 127, 129, 131, 136]; (iii) representing expertise through ontologies and using them to search for experts [114]; and (iv) clustering documents based on their keywords, allowing the retrieval of experts associated to documents in the same clusters with related keywords [130]. Some studies use alternative information sources: (i) bibliographic network information [107, 113]; and (ii) online activities such as posts in CQAs [15, 90], blogs [84], and SNSs [108]. There are also specific approaches, such as converting tag-based classification of questions in CQAs to topic model representations in order to find relevant experts for a given question [32] or using geo-tagged information to locate experts associated with certain places [83].

While most topic model–based approaches design a model to represent each expert, there are proposals where an author can have multiple personas based on the view that the author can write about different combinations of topics for each publication [102]. With the aid of bibliographic network information, central authors (well cited and/or with many coauthorships) can be found. This centrality can be used as an indication of expertise [113]. Citation counts, which are related to the longevity of research topics (same topic present over extended periods of time), also are regarded as an indication of expertise [107]. Some studies combine expertise evidence and social network relationships (e.g., coauthorship or online community metadata) to find experts [38, 84, 88, 90].

Domain-specific approaches, such as examining how *difficult questions* are answered in CQA sites, are also an indication that there is expertise [15]. Machine-learning approaches to locate *future* experts in CQA sites, combining features from several types (textual, behavioral, and time-aware), have been proposed as well [132].

CSSeer [27] locates experts using data available in CiteSeer, supported by data extracted from Wikipedia. First, it extracts key phrases (bi-, tri- and quadgrams) from Wikipedia pages about computer science, statistics, and mathematics. The key phrases that appear at least three times in the collection of documents from CiteSeer are considered *key phrase candidates* and the documents are indexed based on these key phrases. To locate an expert, given a query input (set of words), it locates all authors from documents textually relevant (based on the key phrases), giving higher qualification to those with more documents relevant to the query and higher citation count.

Combining various data sources as expertise input to locate experts is introduced by Pal et al. [111]. They crawled data related to 20,000 IBM employees from various online sources, such as blogs, microblogs, wikis, forums, and online profiles. Several features are introduced in the proposed framework. First, they filter the data using an ngram-classifier to select only documents written in English. LDA is applied to calculate the topics of each document. Using a question modeler (linear SVM classifier) and a self-developed algorithm (DOCSENSE), they extract several features to classify the documents. These include content features (topic distribution, hashtags, referenced entities, etc.), social features (is it a reply for a question, a recommendation, something being shared, etc.), processed features (DOCSENSE features such as whether the document is non-relevant for expertise analysis, whether it is a duplicate from other document, etc.) and, lastly, reply features (relating different documents, such as question and reply in forums).

To index the documents, Apache Lucene was used. Each kind of source (forum, blog, etc.) is indexed separately in order to be treated individually. Before retrieving the documents, they apply a query expansion system based on related words identified through the document topics built earlier. They introduce a new relevant score in Lucene (DOCREL), which considers the proximity between query words in the retrieved documents to rank them. Lastly, they use the Gaussian mixture model (GMM) to discard retrieved documents whose topic distribution is not relevant to the query topics. Once the relevant documents are retrieved, the relative expertise score of each document is compared to the retrieved documents from the same source. The expertise score for each source is then calculated and an SVM rank aggregation algorithm combines the various source scores to calculate the final expertise score.

6.2 Expert Ranking

When there is an expertise of interest and several candidates, the goal of expert ranking [37, 48, 98, 121, 125, 131, 144, 147] is to rank these candidates according to their level of expertise or another factor of interest. Most expertise retrieval methods include expert ranking as part of their process since having a ranked list makes more sense than an unordered list [92].

Most studies adopt *graph-based* approaches, such as random walk, to rank retrieved experts [48, 125, 144]. Alternative approaches include (i) applying neural networks in combination with random-walk methods to rank experts [147]; and (ii) using a regularization framework, applied to a heterogeneous network, and comprising authors and documents [37].

Some studies do not adopt a graph-based approach, and the techniques vary. Some use the number of citations from an author's articles as a ranking factor [131]. The voting model, a ranking technique from the area of *data fusion*, is also used [98]. Interested user-centric approaches, such as ranking the experts based on knowledge gain and ease of access, were also found in the literature [121].

Das Gollapalli et al. [48] introduce an expert ranking method based on two techniques. One technique uses a self-developed Author-Document-Topic (ADT) model (a weighted tripartite graph of authors, documents, and topics) while the other is based on PageRank. The ADT model is built on a per-query basis: given a query, the relevant documents are retrieved and introduced in the graph. Their topics are introduced as well (using precalculated associated weights, such as LDA). Lastly, the authors associated to the documents are introduced in the graph as well as nodes representing the initial query from a user. Once the graph is built, three methods are proposed to calculate the "similarity" between the query nodes and a given author: *MaxPath*—the shorter the path, the stronger the similarity; *SumPath*—the more and stronger the paths, the more similar they are; and *ProductPath*—same as SumPath, but multiplies the path weights instead of adding them.

In the PageRank-based approach, given a query, an initial set of documents is retrieved. A graph is built using this documents and their associated authors. Related documents and authors (e.g., through citations) are introduced in this graph. Then a "random surfer" is simulated over this graph and the probability of it reaching a given author node is calculated, establishing the author ranking.

6.3 Expert Profiling

Expert profiling provides a virtual representation of a person based on the individual's expertise [7, 9, 77, 81, 84, 85, 104, 114, 145]. An important factor in expert profiling is deciding which elements/features are important to include in a person's profile [9, 77]. Some of the studies that focus on expert finding generate profiles during their procedures and can be used for expert profiling [7].

A profile is not necessarily human understandable, i.e., it may not be clear which topics/expertise a person has. For example, in a topic model representation, a topic may be just a cluster of words and will be up to the user to deduce the meaning. Some studies rely on external support, such as ontologies, to build human-understandable profiles [114].

There are many approaches to forming expert profiles. They vary both with regard to their techniques and to data sources drawn on: (i) some use online information, such as intranet web pages [151], LOD [77], or topics in online communities data [84]; (ii) other studies refer to social relations to help build the profile, such as propagating expertise [56, 81] or inferring expertise [53] through related authors; and (iii) there are works that analyze expertise in the context of temporal evolution and demonstrate how the expertise evolves over time [33, 43, 85, 104, 116, 145].

Fang et al. [43] introduce an interesting application. It calculates how probable it is that, given an expert, the expert will stay in one's current areas of expertise or will change to new areas. They analyze how the publications associated to an expert vary their topics over time. To define the topics associated to a document, they use the associated key words. All of the abstracts of documents associated to a given key word are then analyzed to define the topic model (set of words) associated with the topic.

Based on the volume of publications associated to each topic in each year, they introduce a probabilistic model to calculate whether an expert (i) will stay in one's current research areas or (ii) will migrate to new areas. To define which path the expert will take, three features are considered: (i) how common is it for the expert to change areas based on past years, (ii) how similar is a new area to the expert's current areas, and (iii) how popular is the new area based on existing publications from other experts.

By treating the topics related to a given expert in a given year as a set, they introduce a predictive language model (PLM) over these sets of topics (represented by the topic model words associated to them) and their associated probabilities, previously calculated. Given a query topic, the PLM calculates how probable it is that an expert will research the topic.

Author2Vec [56] is an unsupervised machine-learning approach to estimate an author's representation as a vector of embeddings extracted from the author's papers using Paragraph2Vec. The distance and angular similarity between vectors (the author vector and learning paper vector) is used by a neural language model to learn the author's vector representation. The neural network is supplied with both positive (documents produced by the author) and negative (documents not produced by the author) input. Given an input, it will output a weight indicating how probable it is for the author to write about it.

6.4 Expert Clustering

Automated expertise retrieval makes it possible to cluster people in terms of their similar expertise. A number of studies use graphs and similarity metrics for this task. The similarity is calculated from the contents of associated documents [19, 28, 79, 148] but might also include documents' metadata, such as publication venue and coauthorship information [56, 130]. Additional information might be added, such as work relationships [10]. Other works cluster experts based on their expertise representation, using techniques such as structural regularity [133].

While most studies concentrate on content-based topic model similarity [28, 79, 148], there are others that examine similar authors who are cited together [130] or have a social proximity based on previous collaboration information [56].

Boeva et al. [19] introduce an expert clustering approach by partitioning experts based on the key words associated to their documents. To extract these key words from the expert's documents, they apply a part-of-speech tagger to the documents' data and extract three types of key words: (i) *adjective nouns*—an adjective followed by a noun; (ii) *multiple noun*—sequence of nouns; and (iii) *single noun*—the remaining nouns. Once all experts' key words are extracted, they are clustered through a semantic similarity metric based on Wordnet. Each expert profile is transformed into a vector, where each dimension represents the percentage of keywords in the expert profile that are present in the given cluster. Once the expert vectors are built, the Euclidean distance is applied to cluster them and identify similar experts.

6.5 Expert Recommendation

Expert recommendation [22, 25, 26, 30, 47, 72, 123, 126, 140–142, 142] (also called *matching* in the literature [141]) is concerned with recommending others to interact with a given expert. Expert recommendation might seek to match experts with similar profiles (similar to clustering) but also with experts who could make a worthwhile collaboration through their complementary expertise. For example, a text-sequence processing expert could collaborate with a DNA mapping expert. This opens up a fruitful topic in expert recommendation: how should one match experts from different domains of knowledge, such as medicine and computing? Few studies [4, 126] have addressed this issue.

The techniques adopted in the literature to carry out expertise recommendation vary. Most use expertise evidence combined with social relations [22, 25, 26, 30, 72, 126, 140–142, 149]. Some studies adopt alternative approaches, such as (i) creating a grant database, associated with content similarity and the collaboration network, to suggest new collaborations between researchers with different areas of expertise [4]; (ii) introducing path optimization to graphs linking authors and article contents, where, for example, a path *Author1-Paper1-Term-Paper2-Author2* becomes an *Author-Term-Author2* path, thus resulting in a smaller graph and improving random-walk algorithm application [149]; and (iii) using concepts from the *expertise seeking* area [54].

Cohen and Ebel [30] introduce a researcher collaboration suggestion based on the researcher's social network (collaborations) and a given topic of interest for collaboration, defined by key words. A graph is built that includes the authors (vertices where they are represented by a bag of words from their publication titles) and collaborations between authors (edges composed by three features: the publication title, date, and venue). Over the graph, a query composed by an author (a vertex) and a set of key words is executed.

Score functions are applied to determine how probable it is that a given vertex (author) will collaborate with another vertex. The first function calculates the structural proximity, which has two basic approaches: one uses the distance between nodes weighted by a given function; the other calculates the structural proximity based on the common collaborations between the vertices (i.e., past collaborations).

The second function calculates the textual relevancy, i.e., how probable is it that a given expert will work on the topic specified by the key words from the query. Two approaches are introduced. The first uses TF-IDF between the expert profile key words and the query key words. The second approach uses a self-developed function called *Collab*. *Collab* considers the previous collaborations of a given expert to determine whether the expert is relevant to the query. For each previous collaboration, it calculates how relevant it is to the query (TF-IDF on the key words and collaboration title), how much time has passed since the collaboration (logarithm function), and whether it occurred in a venue where the query expert has already published. In its last step, *Collab* sums the previous values from all of the neighbors of a given node to calculate the weight (relevance) of the node to the original query node.

Lastly, the authors combine both structural proximity and textual relevancy through a *CScore* function, which does a weighted sum of the scores from the previous introduced functions.

7 CURRENT WORK OVERVIEW

In this section, a comparison is made between a selected set of works related to expertise retrieval so that they can be classified in the proposed taxonomy. The selection was made on the basis of two criteria: first, to cover examples in the full range of taxonomic classifications; and, second, to include the most recent or relevant ones. This resulted in 26 works, which are compared in Tables 2, 3, and 4. The tables are underpinned by the four components that guide our taxonomy: data source, data extraction, expertise extraction, and application (through its single facet, *task*). The following types of behavior were observed in the current studies that were surveyed:

- Most works use public unrestricted, plain-text sources, such as public sites and other data sources that have no clearly defined structure. Since the most common expertise extraction methods are based on specific terms (such as topic models, language models, term frequency), structured data is not a requirement.
- A complete data-extraction process is more common than a focused data extraction given the fact that many works form an expert representation in advance, thus, allowing browsing and searching for the extracted expertise information. With regard to expert composition, there is no clear predominance between simple and complex approaches.
- Term-based expertise extraction (topic models, language models, and term frequency) can be found in many works. Graph-based expertise extraction is also common, especially for complex expert composition. This is natural since it is a good way of designing relations such as those between author and venue and author and coauthor as well as finding citations on documents and/or people.

Table 2. Current Work Comparison - 1/3

Work	Data source	Data extraction	Expertise extraction	Task (Application)
Li2015B [80]	Private unstructured plain text	General simple expert composition	Custom temporal continuous with semantic support	Expert profiling (builds topic cloud views of expertise models based on CVs)
Xu2012 [140]	Public unrestricted, unstructured plaintext	General complex expert composition	Graph and custom, with lexical database support	Expert recommendation (suggests collaborators based on scientific publications)
Chen2013 [27]	Public unrestricted, semistructured dataset	General simple expert composition	Term frequency extraction, encyclopedia semantic support	Expert finding (finds experts based on CiteSeer and Wikipedia data)
Fang2014 [43]	Public unrestricted, unstructured plain text	General complex expert composition, with stemming and stop words removal	Language model extraction, time slice temporal support	Expert profiling (based on previous publications, analyzes whether an author may change a research line in the future)
Parada2013 [112]	Private semistructured and structured dataset	General complex expert composition	Graph extraction	Expert recommendation (uses a social-based calculated PCI [Potencial Collaboration Index] to recommend collaborations)
Pal2015 [111]	Public unrestricted unstructured plain text	Focused simple expert composition	Topic model extraction	Expert finding (combines multiple data sources to locate an expert)
Gysel2016 [134]	Public unrestricted unstructured plain text	General simple expert composition with stop words	Term frequency extraction	Expert finding (promotes expert finding optimization using back-propagation neural networks)
Rybak2014 [116]	Public restricted semistructured dataset	General simple expert composition	Custom extraction, with ontology semantic and temporal support	Expert profiling (Uses an ontology to show how the expertise of an author changes over time)
Liu2013B [91]	Public unrestricted, unstructured communications	Focused complex expert composition, stop words removal	Graph and term frequency extraction	Expert finding (Analyzes interactions in a Q&A site to locate experts)
Gollapalli2013 [48]	Public unrestricted, semistructured dataset and plain text	Focused simple expert composition	Graph and topic model extraction	Expert ranking (gathers Arnetminer data, applying a modified PageRank and a tripartite graph algorithm)
Boeva2014 [19]	Public unrestricted, unstructured plaintext	General simple expert composition, with stemming and stop words removal	Topic model extraction	Expert clustering (through the author profiles, keywords, compared using Wordnet)

- Most of the studies focus on expert finding, followed by those focused on expert recommendation. This indicates how finding experts on a given topic is a major factor in expertise retrieval.

8 CONCLUSION AND OPEN ISSUES

Automated expertise retrieval is a process that can be used in many applications, such as expert finding and expert profiling. In this article, we created a taxonomy that covers many of existing schemes. It is structured in four components: (i) data source, (ii) data extraction, (iii) expertise extraction, and (iv) application. Using the taxonomy as a guideline, we introduced several factors related to expertise retrieval and described some of them in considerable detail. Through our analysis, we were able to identify some open issues from the expertise retrieval task. In this final section, we summarize these issues while making recommendations for future research.

Table 3. Current Work Comparison - 2/3

	Data source	Data extraction	Expertise extraction	Task (Application)
Kaya2014 [69]	Public unrestricted, semistructured dataset	General complex expert composition	Term frequency extraction, with time slice temporal support	Expert finding and profiling (builds a data cube based on the publication data and applies OLAP methods to locate and profile experts)
Chaiwanarom2015 [25]	Public restricted semistructured dataset	General complex expert composition	Graph and topic model extraction, with time slice temporal support	Expert matching (suggests potencial collaborations based on social relations, researcher seniority and publications' content similarity)
Fang2010 [44]	Public unrestricted, unstructured plain text	Focused simple expert composition, with stemming	Custom and language model extraction	Expert finding (introduces a discriminative model to associate authors to documents)
GaneshJ2016 [56]	Public unrestricted semistructured and unstructured dataset and plain text	General simple expert composition	Custom extraction	Expert profiling (uses a neural network to learn how to associate authors to documents)
Mangaravite2016a [99]	Public unrestricted, semistructured dataset	General simple expert composition, with stop words removal	Language model extraction	Expert finding (introduces new normalization techniques to weights associating authors and documents)
Yang2015 [142]	Private semistructured, unstructured dataset, plain text	General complex expert composition	Graph and language model extraction	Expert recommendation (suggests potential collaborations based on publications' content similarity and relations in a scientific social network)
Liu2014A [92]	Public unrestricted, unstructured plain text	Focused complex expert composition	Language model extraction	Expert finding (introduces AMiner-mini, a version of Arnetminer applicable in institutions)
Neshati2014 [107]	Public unrestricted, unstructured plain text	Focused simple expert composition	Language model extraction	Expert finding (locates leading authors in a publication)
Balog2009 [7]	Public unrestricted, unstructured plain text	General and focused expert composition, with stop words removal	Language model extraction	Expert finding (introduces a language modeling framework for expert finding)

8.1 Expertise Association

All methods of automated expertise retrieval must address certain issues related to the person versus expertise association. Some issues that we regard as important for a good expertise retrieval are listed below.

(1) How can one associate a person with a document? When there is no clear person-document association, how can one find this out?
(2) If a document is associated with more than one person, who is responsible for each item of expertise evidence?
(3) How important and/or reliable is a document as a means of representing the expertise?

In our view (corroborated by other research studies [11]), these are not trivial issues. With regard to the task of associating people with evidence, the studies in the literature vary a good deal. Some use metadata (such as bibliographic networks, post authors in social networks/forums or e-mail header information) [8, 13, 87, 117, 121] while others use the person's name and/or e-mail address in the document, which can cause problems such as ambiguity in the name and other issues [6, 9, 44, 107]. Finding reliable ways to associate people and evidence is still an open topic of research.

Table 4. Current Work Comparison - 3/3

	Data source	Data extraction	Expertise extraction	Task (Application)
Cohen2013 [30]	Public unrestricted semistructured dataset	General complex expert composition	Graph and term frequency extraction with time slice temporal support	Expert recommendation (uses a start researchers, key words and coauthoring network to suggest collaborations)
Zhu2014 [150]	Public unrestricted, unstructured plain text	General simple expert composition	Topic model extraction	Expert finding (besides experts in the desired area, also includes experts on other related areas)
Deng2012 [37]	Public unrestricted, semistructured dataset	Focused complex expert composition	Graph extraction	Expert ranking (using coauthoring network and citations)
Li2015 [79]	Public unrestricted unstructured plaintext	General simple expert composition	Topic model extraction	Expert clustering (analyzes publications' content similarity)
Kumar2016 [75]	Public unrestricted, unstructured plain text	General complex expert composition	Term frequency extraction	Expert finding (through Q&A site data, considering *best-answers* indications as expertise hints)
Osborne2013 [110]	Public unrestricted, semistructured and unstructured dataset and plain text	General complex expert composition	Custom extraction	Expert finding and profiling (introduces the Rexplore tool to visualize and relate author publications and expertise)

With regard to the question of linking expertise evidence to people, one approach is to consider the proximity between an item of evidence and a person's name in a given document. The closer the name is to the evidence, the greater is the chance that the evidence represents the person's expertise [2]. This method can be effective in some domains, such as web pages, for example. However, it is not applicable in scientific papers, where the author's name usually appears on the first page and not near the expertise evidence in the document [107].

The issue of determining in a multiple-author document how well each author expertise is represented is still an open issue. Preliminary work on the topic has been done using, for example, the order of authors in publications as an indication of expertise degree. In this case, the first authors are considered the main contributors for a given paper [82, 97].

The methods employed to establish how important and/or reliable the evidence is to estimate a person's expertise in a given document varies in accordance with the document type. In web pages, the number of referral links can be treated as a measure of reliability (*Page Rank*) [152]. In the case of scientific articles, the number of citations can be used [27]. Depending on their importance, some studies analyze the author publications over a period of time and seek to identify the preferred topics [33, 43, 116].

8.2 Combining Multiple Evidence

A wide range of features is taken into account when analyzing expertise evidence, such as topic models, social relations, and semantic analysis through Wikipedia articles. However, only a few of them [31, 44] address the question of how to combine these features to improve the results. Learning approaches, such as neural networks based on user feedback, could provide new and useful ways to combine expertise evidence.

Clearly, in different domains and applications, the importance of each type of evidence may vary owing to the quality of the data used for expertise evidence. For example, in the scientific domain, posts in a social network should weight lower than a paper published in the proceedings of a prestigious conference.

8.3 Multiple Languages

To the best of our knowledge, none of the existing studies takes account of the fact that expertise evidence can be in more than one language. In our view, in several cases, it is an advantage to be able to correlate the same expertise described in several languages. An attempt is being made by one researcher to relate knowledge from the same domain in different languages [5], but it is still in its early stages.

The support of multiple languages would benefit the expertise retrieval process in several situations. A multinational company, which has documents in several languages, could improve interaction between teams by drawing on shared expertise. Localized research results, published in different languages (such as Chinese or Hindi), could serve as expertise evidence for locating research experts. Researchers can publish a new work (with initial results) first for a local event in their own language and later on at an international conference in English.

We suggest using a common semantic mediator, such as Wikipedia, for this task. Wikipedia has been successfully used [27, 35] for semantic analysis. As it provides pages in several languages, and identifies which pages correspond to the same concept, a method could be devised to identify related expertise evidence between languages.

8.4 Data Veracity

Another open issue that should be pointed out is how to assess the quality/trust of expertise evidence. Analyzing data veracity is a common problem in big data integration [40]. In the context of expertise evidence, we have put forward some features that are found in expertise retrieval that could be combined with an analysis of data veracity:

- The recognition of the conference or journal where an article was published can indicate its quality. To assess this, one could consider the citation count of the published articles, for example.
- An article published in a conference proceedings, where the members of the Program Committee have expertise over the topics contemplated by the article, has a greater chance of providing better standards of expertise evidence.
- The impact level of scientific publications as measured by established metrics, such as H-Index [52] or JCR [3], is also a strong indicator of quality/trust.
- Good evidence tends to be well cited. Thus, it could be worth investigating how well referenced a given item of expertise evidence is.
- The reputation of a site (*url*), where a page is published, also serves to ensure its data veracity.
- User feedback can help in classifying evidence, for example, users *rating* documents in scientific indexers or event sites.

8.5 User Interaction

There are many methods that can be employed for automated expertise retrieval. However, to the best our knowledge, there is none that takes full account of user feedback during the process. Some do so in a limited way, for example, by classifying documents [51] or expert matching [129] as relevant, irrelevant, or false. Our study corroborates the findings of Balog et al. [11], who reached the same conclusion.

When including user feedback, we suggest defining the intention of the user when working with expertise retrieval. This could be done by measuring the user's degree of satisfaction with the current results through some metric while ensuring that there is a minimum amount of interaction during the process. User interaction can be regarded not only as a way to adjust parameters but also as a way to make alterations in design decisions on automated expertise retrieval and related tasks. This could ensure a more general, interchangeable, and component-based approach that is easily adapted to new domains and data formats based on user feedback.

8.6 Explanation of the Results

Helping the user understand the results is another topic that we found to be ignored in the literature. The better the user understands the results, the more confidence the user will have in the expertise retrieval system. A system could be adopted to help the user understand this by, for example, describing how a given expertise was captured and how the system assesses its relevance.

The current approaches usually give a list of people or a graph cluster as a result but fail to describe how they obtained this data. Adopting an approachin which the user does not understand how the result was achieved is not appropriate when dealing with people. Collaborations based on false assumptions could, for example, result in unsuccessful social interactions and, thus, should be avoided. By allowing the user to understand the result, the user is free to use one's own judgment and decide whether to go ahead with contacting the referred expert.

8.7 Description of the Expertise

As in the case of the explanation of results, the description of expertise is another open issue. This category can be defined as a clear, concise, and preferably human-readable view of a person's expertise.

To the best of our knowledge, there are few approaches to automatically making a human-readable representation of expertise [9, 77]. This kind of representation could be used to assess a system's quality in providing one's expertise by comparing it with the expertise obtained from several systems. This would naturally require a common representation for expertise so that a comparison could be made.

8.8 Contextual Analysis

Research and expertise do not evolve per se, but rather as a result of events in the context of the involved people or topic. For example, researchers in academia may start working on a new topic based on a ground-breaking article [137]. Professionals in the industry may change their expertise interests as the result of a significant recent event. An awareness of context when analyzing expertise may yield interesting results and assist in understanding the evolution of expertise and track changes in the topics of interest over a period of time.

8.9 Cross-Domain Collaboration

Another topic that has not received attention is expertise collaboration between different domains, for example, between biology and computer science. Cross-domain collaboration is more difficult than intradomain collaboration since it is not a trivial task to identify the related work between different domains. For example, research on string similarity and substring analysis from computer science can be applied to DNA sequencing; but how can we find such a relationship? Some approaches draw on studies in the literature [126] to find possible collaborations.

We suggest using a semantic mediator to identify conceptual relationships and find possible new forms of collaborations. Wikipedia is one example of a possible semantic mediator. A system can be devised where people input their problem description and related techniques and research could be suggested.

8.10 Implementation and Information Exchange

Few studies [50, 73, 127, 134] have analyzed questions related to implementation or scalability when introducing their schemes. Automated expertise retrieval can be regarded as a problem in the domain of BDI (Big Data Integration) [40] since it can handle a lot of data and must integrate and order this data to extract an item of expertise evidence and representation. Thus, issues such as how expertise representation should be indexed and its searches facilitated are open to further suggestions and improvements.

A standard expertise representation, which could be exchanged between systems, is also another interesting area of research. In the context of big data, where expertise evidence is retrieved by several systems that can exchange this information later, could provide a useful way to scale systems and compare different approaches to determine which are best for a given domain.

REFERENCES

[1] Serge Abiteboul, Peter Buneman, and Dan Suciu. 2000. *Data on the Web: From Relations to Semistructured Data and XML*. Morgan Kaufmann Publishers Inc., San Francisco, CA.

[2] F. Alarfaj, U. Kruschwitz, and C. Fox. 2013. An adaptive window-size approach for expert-finding. In *CEUR Workshop Proceedings*, Vol. 986. Delft University of Technology, The Netherlands and Centrum Wiskunde en Informatica Amsterdam, The Netherland, 76–79.

[3] Clarivate Analytics. 2017. Journal Citation Reports. Retrieved February 2, 2017 from http://about.jcr.incites.thomsonreuters.com/.

[4] Masataka Araki, Marie Katsurai, Ikki Ohmukai, and Hideaki TAkeda. 2017. Interdisciplinary collaborator recommendation based on research content similarity. *IEICE Transactions on Information and Systems* 100, 4, 1–8. DOI: https://doi.org/10.1587/transinf.E100.D.1

[5] Liangming Pan B, Zhigang Wang, Juanzi Li, and Jie Tang. 2011. Domain specific cross-lingual knowledge linking based on similarity flooding. 7091, 426–438. DOI: https://doi.org/10.1007/978-3-642-25975-3

[6] Krisztian Balog, Leif Azzopardi, and Maarten de Rijke. 2006. Formal models for expert finding in enterprise corpora. In *Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'06)*. ACM, New York, NY, 43–50. DOI: https://doi.org/10.1145/1148170.1148181

[7] Krisztian Balog, Leif Azzopardi, and Maarten de Rijke. 2009. A language modeling framework for expert finding. *Information Processing & Management* 45, 1, 1–19. DOI: https://doi.org/10.1016/j.ipm.2008.06.003

[8] Krisztian Balog and Maarten de Rijke. 2006. Finding experts and their eetails in e-mail corpora. In *Proceedings of the 15th International Conference on World Wide Web (WWW'06)*. ACM, New York, NY, 1035–1036. DOI: https://doi.org/10.1145/1135777.1136002

[9] Krisztian Balog and Maarten De Rijke. 2007. Determining expert profiles (with an application to expert finding). In *Proceedings of the 20th International Joint Conference on Artificial Intelligence (IJCAI'07)*. Morgan Kaufmann Publishers Inc., San Francisco, CA, 2657–2662. http://dl.acm.org/citation.cfm?id=1625275.1625703.

[10] Krisztian Balog and Maarten de Rijke. 2007. Finding similar experts. In *Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'07)*. ACM, New York, NY, USA, 821–822. DOI: https://doi.org/10.1145/1277741.1277926

[11] Krisztian Balog, Yi Fang, Maarten de Rijke, Pavel Serdyukov, and Luo Si. 2012. Expertise retrieval. *Foundations and Trends® in Information Retrieval* 6, 2–3, 127–256. DOI: https://doi.org/10.1561/1500000024

[12] Z. Ban and L. Liu. 2016. CICPV: A new academic expert search model. In *IEEE 30th International Conference on Advanced Information Networking and Applications (AINA'16)*. IEEE, 47–52. DOI: https://doi.org/10.1109/AINA.2016.14

[13] Richard Berendsen, Maarten de Rijke, Krisztian Balog, Toine Bogers, and Antal van den Bosch. 2013. On the assessment of expertise profiles. *Journal of the American Society for Information Science and Technology* 64, 10, 2024–2044. DOI: https://doi.org/10.1002/asi.22908

[14] Michael K. Bergman. 2001. White paper: The deep web: surfacing hidden value. *The Journal of Electronic Publishing* 7, 1, online. Retrieved July 31, 2019 from http://dx.doi.org/10.3998/3336451.0007.104

[15] M. Bhanu and J. Chandra. 2016. Exploiting response patterns for identifying topical experts in stackoverflow. In *11th International Conference on Digital Information Management (ICDIM'16)*. 139–144. DOI: https://doi.org/10.1109/ICDIM.2016.7829790

[16] Christian Bizer, Tom Heath, Kingsley Idehen, and Tim Berners-Lee. 2008. Linked data on the web (LDOW'08). In *Proceedings of the 17th International Conference on World Wide Web (WWW'08)*. ACM, New York, NY, 1265–1266. DOI: https://doi.org/10.1145/1367497.1367760

[17] David M. Blei. 2012. Probabilistic topic models. *Communications of the ACM* 55, 4, 77–84. DOI: https://doi.org/10.1145/2133806.2133826

[18] David M. Blei, Andrew Y. Ng, and Michael I. Jordan. 2003. Latent Dirichlet allocation. *J. Mach. Learn. Res.* 3, 993–1022. http://dl.acm.org/citation.cfm?id=944919.944937.

[19] Veselka Boeva, Liliana Boneva, and Elena Tsiporkova. 2014. *Semantic-Aware Expert Partitioning*. Springer International Publishing, Cham, 13–24. DOI: https://doi.org/10.1007/978-3-319-10554-3_2

[20] Levent Bolelli, Şeyda Ertekin, and C. Lee Giles. 2009. Topic and trend detection in text collections using latent Dirichlet allocation. In *Proceedings of the 31st European Conference on IR Research on Advances in Information Retrieval.* Springer, Berlin, 776–780. DOI: https://doi.org/10.1007/978-3-642-00958-7_84

[21] S. Budalakoti, D. DeAngelis, and K. S. Barber. 2009. Expertise modeling and recommendation in online question and answer forums. 4, 481–488. DOI: https://doi.org/10.1109/CSE.2009.62

[22] Guillaume Cabanac. 2011. Accuracy of inter-researcher similarity measures based on topical and social clues. *Scientometrics* 87, 3, 597–620. DOI: https://doi.org/10.1007/s11192-011-0358-1

[23] Christopher S. Campbell, Paul P. Maglio, Alex Cozzi, and Byron Dom. 2003. Expertise identification using email communications. In *Proceedings of the 12th International Conference on Information and Knowledge Management (CIKM'03).* ACM, New York, NY, 528–531. DOI: https://doi.org/10.1145/956863.956965

[24] Youngchul Cha, Keng-hao Chang, Hari Bommaganti, Ye Chen, Tak Yan, Bin Bi, and Junghoo Cho. 2015. A universal topic framework (UniZ) and its application in online search. In *Proceedings of the 30th Annual ACM Symposium on Applied Computing (SAC'15).* ACM, New York, NY, 1078–1085. DOI: https://doi.org/10.1145/2695664.2695806

[25] Paweena Chaiwanarom and Chidchanok Lursinsap. 2015. Collaborator recommendation in interdisciplinary computer science using degrees of collaborative forces, temporal evolution of research interest, and comparative seniority status. *Knowledge-Based Systems* 75, 161–172. DOI: https://doi.org/10.1016/j.knosys.2014.11.029

[26] Hung-Hsuan Chen, Liang Gou, Xiaolong Zhang, and Clyde Lee Giles. 2011. CollabSeer: A search engine for collaboration discovery. In *Proceedings of the 11th Annual International ACM/IEEE Joint Conference on Digital Libraries (JCDL'11).* ACM, New York, NY, 231–240. DOI: https://doi.org/10.1145/1998076.1998121

[27] Hung-Hsuan Chen, Pucktada Treeratpituk, Prasenjit Mitra, and C. Lee Giles. 2013. CSSeer: An expert recommendation system based on CiteSeerX. In *Proceedings of the 13th ACM/IEEE-CS Joint Conference on Digital Libraries (JCDL'13).* ACM, New York, NY, 381–382. DOI: https://doi.org/10.1145/2467696.2467750

[28] Xu Chen, Mingyuan Zhou, and Lawrence Carin. 2012. The contextual focused topic model. In *Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD'12).* ACM, New York, NY, 96–104. DOI: https://doi.org/10.1145/2339530.2339549

[29] X. Cheng, S. Zhu, G. Chen, and S. Su. 2015. Exploiting user feedback for expert finding in community question answering. In *IEEE International Conference on Data Mining Workshop (ICDMW'15).* IEEE, 295–302. DOI: https://doi.org/10.1109/ICDMW.2015.181

[30] Sara Cohen and Lior Ebel. 2013. Recommending collaborators using keywords. In *Proceedings of the 22nd International Conference on World Wide Web (WWW'13 Companion).* ACM, New York, NY, 959–962. DOI: https://doi.org/10.1145/2487788.2488091

[31] Ronan Cummins, Mounia Lalmas, and Colm O'Riordan. 2010. Learning aggregation functions for expert search. In *Proceedings of the 19th European Conference on Artificial Intelligence (ECAI'10).* IOS Press, Amsterdam, the Netherlands, 535–540. http://dl.acm.org/citation.cfm?id=1860967.1861072.

[32] Arash Dargahi Nobari, Sajad Sotudeh Gharebagh, and Mahmood Neshati. 2017. Skill Translation models in expert finding. *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'17).* 1057–1060. DOI: https://doi.org/10.1145/3077136.3080719

[33] Ali Daud. 2012. Using time topic modeling for semantics-based dynamic research interest finding. *Knowledge-Based Systems* 26, 154–163. DOI: https://doi.org/10.1016/j.knosys.2011.07.015

[34] Ali Daud, Juanzi Li, Lizhu Zhou, and Faqir Muhammad. 2009. Exploiting temporal authors interests via temporal-author-topic modeling. In *Proceedings of the 5th International Conference on Advanced Data Mining and Applications (ADMA'09).* Springer, Berlin, 435–443. DOI: https://doi.org/10.1007/978-3-642-03348-3_42

[35] Elnaz Davoodi, Keivan Kianmehr, and Mohsen Afsharchi. 2013. A semantic social network-based expert recommender system. *Applied Intelligence* 39, 1, 1–13. DOI: https://doi.org/10.1007/s10489-012-0389-1

[36] B. de La Robertie, Y. Pitarch, A. Takasu, and O. Teste. 2017. Identifying authoritative researchers in digital libraries using external a priori knowledge. In *Proceedings of the Symposium on Applied Computing (SAC'17).* ACM, New York, NY, 1017–1022. DOI: https://doi.org/10.1145/3019612.3019809

[37] Hongbo Deng, Jiawei Han, Michael R. Lyu, and Irwin King. 2012. Modeling and exploiting heterogeneous bibliographic networks for expertise ranking. In *Proceedings of the 12th ACM/IEEE-CS Joint Conference on Digital Libraries (JCDL'12).* ACM, New York, NY, 71–80. DOI: https://doi.org/10.1145/2232817.2232833

[38] H. Deng, I. King, and M. R. Lyu. 2012. Enhanced models for expertise retrieval using community-aware strategies. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)* 42, 1, 93–106. DOI: https://doi.org/10.1109/TSMCB.2011.2161980

[39] Byron Dom, Iris Eiron, Alex Cozzi, and Yi Zhang. 2003. Graph-based ranking algorithms for e-mail expertise analysis. In *Proceedings of the ACM SIGMOD Workshop on Research Issues in Data Mining and Knowledge Discovery*. ACM, 42–48, DOI: https://doi.org/10.1145/882082.882093

[40] X. L. Dong and D. Srivastava. 2013. Big data integration. In *IEEE 29th International Conference on Data Engineering (ICDE'13)*. IEEE, 1245–1248. DOI: https://doi.org/10.1109/ICDE.2013.6544914

[41] Jianguang Du, Jing Jiang, Dandan Song, and Lejian Liao. 2015. Topic Modeling with Document Relative Similarities. In *Proceedings of the 24th International Conference on Artificial Intelligence (IJCAI'15)*. AAAI Press, 3469–3475. http://dl.acm.org/citation.cfm?id=2832581.2832732.

[42] Trong Hai Duong, Ngoc Thanh Nguyen, and Geun Sik Jo. 2010. Constructing and mining a semantic-based academic social network. *J. Intell. Fuzzy Syst.* 21, 3, 197–207. http://dl.acm.org/citation.cfm?id=1735086.1735091.

[43] Yi Fang and Archana Godavarthy. 2014. Modeling the dynamics of personal expertise. In *Proceedings of the 37th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'14)*. ACM, New York, NY, 1107–1110. DOI: https://doi.org/10.1145/2600428.2609521

[44] Yi Fang, Luo Si, and Aditya P. Mathur. 2010. Discriminative models of integrating document evidence and document-candidate associations for expert search. In *Proceedings of the 33rd International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'10)*. ACM, New York, NY, 683–690. DOI: https://doi.org/10.1145/1835449.1835563

[45] Maryam Fazel-Zarandi and Mark S. Fox. 2011. Constructing expert profiles over time for skills management and expert finding. In *Proceedings of the 11th International Conference on Knowledge Management and Knowledge Technologies (i-KNOW'11)*. ACM, New York, NY, Article 5, 6 pages. DOI: https://doi.org/10.1145/2024288.2024295

[46] Shengxiang Gao, Xian Li, Zhengtao Yu, Yu Qin, and Yang Zhang. 2017. Combining paper cooperative network and topic model for expert topic analysis and extraction. *Neurocomputing* 257, 136–143. DOI: https://doi.org/10.1016/j.neucom.2016.12.074 Machine Learning and Signal Processing for Big Multimedia Analysis.

[47] Sujatha Das Gollapalli, Prasenjit Mitra, and C. Lee Giles. 2012. Similar researcher search in academic environments. In *Proceedings of the 12th ACM/IEEE-CS Joint Conference on Digital Libraries (JCDL'12)*. ACM, New York, NY, 167–170. DOI: https://doi.org/10.1145/2232817.2232849

[48] Sujatha Das Gollapalli, Prasenjit Mitra, and C. Lee Giles. 2013. Ranking experts using author-document-topic graphs. In *Proceedings of the 13th ACM/IEEE-CS Joint Conference on Digital Libraries (JCDL'13)*. ACM, New York, NY, 87–96. DOI: https://doi.org/10.1145/2467696.2467707

[49] Thomas L. Griffiths and Mark Steyvers. 2004. Finding scientific topics. *Proceedings of the National Academy of Sciences* 101, suppl 1, 5228–5235. DOI: https://doi.org/10.1073/pnas.0307752101

[50] Qi He, Bi Chen, Jian Pei, Baojun Qiu, Prasenjit Mitra, and Lee Giles. 2009. Detecting topic evolution in scientific literature: how can citations help?. In *Proceedings of the 18th ACM Conference on Information and Knowledge Management (CIKM'09)*. ACM, New York, NY, 957–966. DOI: https://doi.org/10.1145/1645953.1646076

[51] Djoerd Hiemstra. 2001. *Using Language Models for Information Retrieval*. Retrieved July 31, 2019 from http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.66.2128&rep=rep1&type=pdf.

[52] J. E. Hirsch. 2005. An index to quantify an individual's scientific research output. *Proceedings of the National Academy of Sciences of the United States of America* 102, 46, 16569–16572. DOI: https://doi.org/10.1073/pnas.0507655102 arXiv:http://www.pnas.org/content/102/46/16569.full.pdf.

[53] Nguyen Le Hoang, Pham Vu Dang Khoa, and Do Phuc. 2013. Predicting preferred topics of authors based on co-authorship network. In *RIVF International Conference on Computing Communication Technologies - Research, Innovation, and Vision for Future (RIVF'13)*. 70–75. DOI: https://doi.org/10.1109/RIVF.2013.6719869

[54] Katja Hofmann, Krisztian Balog, Toine Bogers, and Maarten de Rijke. 2010. Contextual factors for finding similar experts. *Journal of the American Society for Information Science and Technology* 61, 5, 994–1014. DOI: https://doi.org/10.1002/asi.21292

[55] S. Huang, Y. Tang, F. Tang, and J. Li. 2014. Link prediction based on time-varied weight in co-authorship network. In *Proceedings of the IEEE 18th International Conference on Computer Supported Cooperative Work in Design (CSCWD'14)*. IEEE, 706–709. DOI: https://doi.org/10.1109/CSCWD.2014.6846931

[56] Ganesh J, Soumyajit Ganguly, Manish Gupta, Vasudeva Varma, and Vikram Pudi. 2016. Author2Vec: Learning author representations by combining content and link information. In *Proceedings of the WWW - International Conference on World Wide Web*. International World Wide Web Conferences Steering Committee, 49–50. DOI: https://doi.org/10.1145/2872518.2889382

[57] Shoaib Jameel and Wai Lam. 2013. *An N-Gram Topic Model for Time-Stamped Documents*. Springer, Berlin, 292–304. DOI: https://doi.org/10.1007/978-3-642-36973-5_25

[58] Shoaib Jameel and Wai Lam. 2013. An unsupervised topic segmentation model incorporating word order. In *Proceedings of the 36th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'13)*. ACM, New York, NY, 203–212. DOI: https://doi.org/10.1145/2484028.2484062

[59] Y. Jiang, X. Li, and W. Meng. 2014. DiscWord: Learning discriminative topics. In *IEEE/WIC/ACM International Joint Conferences on Web Intelligence (WI'14) and Intelligent Agent Technologies (IAT'14)*, Vol. 2. 63–70. DOI: https://doi.org/10.1109/WI-IAT.2014.81

[60] Jian Jin, Qian Geng, Qian Zhao, and Lixue Zhang. 2017. Integrating the trend of research interest for reviewer assignment. In *Proceedings of the 26th International Conference on World Wide Web Companion (WWW'17 Companion)*. International World Wide Web Conferences Steering Committee, Republic and Canton of Geneva, Switzerland, 1233–1241. DOI: https://doi.org/10.1145/3041021.3053053

[61] Yookyung Jo, John E. Hopcroft, and Carl Lagoze. 2011. The web of topics: Discovering the topology of topic evolution in a corpus. In *Proceedings of the 20th International Conference on World Wide Web (WWW'11)*. ACM, New York, NY, 257–266. DOI: https://doi.org/10.1145/1963405.1963444

[62] Nikhil Johri, Dan Roth, and Yuancheng Tu. 2010. Experts' retrieval with multiword-enhanced author topic model. In *Proceedings of the NAACL HLT Workshop on Semantic Search (SS'10)*. Association for Computational Linguistics, Stroudsburg, PA, 10–18. http://dl.acm.org/citation.cfm?id=1867767.1867769.

[63] Nawarat Kamsiang and Twittie Senivongse. 2014. *An Ontology-Based Methodology for Building and Matching Researchers' Profiles*. Springer, Dordrecht, 455–468. DOI: https://doi.org/10.1007/978-94-007-6818-5_32

[64] Maryam Karimzadehgan, Ryen W. White, and Matthew Richardson. 2009. Enhancing expert finding using organizational hierarchies. In *Proceedings of the 31st European Conference on IR Research on Advances in Information Retrieval (ECIR'09)*. Springer, Berlin,177–188. DOI: https://doi.org/10.1007/978-3-642-00958-7_18

[65] Saurabh Kataria, Prasenjit Mitra, Cornelia Caragea, and C. Lee Giles. 2011. Context sensitive topic models for author influence in document networks. In *Proceedings of the 22nd International Joint Conference on Artificial Intelligence - Volume Three (IJCAI'11)*. AAAI Press, 2274–2280. DOI: https://doi.org/10.5591/978-1-57735-516-8/IJCAI11-379

[66] Noriaki Kawamae. 2010. Author interest topic model. In *Proceedings of the 33rd International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'10)*. ACM, New York, NY, 887–888. DOI: https://doi.org/10.1145/1835449.1835666

[67] Noriaki Kawamae. 2010. Latent interest-topic model: Finding the causal relationships behind dyadic data. In *Proceedings of the CIKM - ACM International Conference on Information and Knowledge Management*. 649–658. DOI: https://doi.org/10.1145/1871437.1871521

[68] Noriaki Kawamae. 2012. Theme chronicle model: Chronicle consists of timestamp and topical words over each theme. In *Proceedings of the 21st ACM International Conference on Information and Knowledge Management (CIKM'12)*. ACM, New York, NY, 2065–2069. DOI: https://doi.org/10.1145/2396761.2398573

[69] Mehmet Kaya and Reda Alhajj. 2014. Development of multidimensional academic information networks with a novel data cube based modeling method. *Information Sciences* 265, 211–224. DOI: https://doi.org/10.1016/j.ins.2013.11.012

[70] Yun Sing Koh and Gillian Dobbie. 2012. Indirect weighted association rules mining for academic network collaboration recommendations. In *Proceedings of the 10th Australasian Data Mining Conference - Volume 134 (AusDM'12)*. Australian Computer Society, Inc., Darlinghurst, Australia, 167–173. DOI: http://dl.acm.org/citation.cfm?id=2525373.2525393

[71] Xiangjie Kong, Huizhen Jiang, Teshome Megersa Bekele, Wei Wang, and Zhenzhen Xu. 2017. Random walk-based beneficial collaborators recommendation exploiting dynamic research interests and academic influence. In *Proceedings of the 26th International Conference on World Wide Web Companion (WWW'17 Companion)*. International World Wide Web Conferences Steering Committee, Republic and Canton of Geneva, Switzerland, 1371–1377. DOI: https://doi.org/10.1145/3041021.3051154

[72] Xiangjie Kong, Huizhen Jiang, Zhuo Yang, Zhenzhen Xu, Feng Xia, and Amr Tolba. 2016. Exploiting publication contents and collaboration networks for collaborator recommendation. *PLOS One* 11, 2, 1–13. DOI: https://doi.org/10.1371/journal.pone.0148492

[73] Yue Kou, Derong Shen, Hongbin Xu, Menger Lin, Ge Yu, and Tiezheng Nie. 2015. Two-level interactive identification and derivation of topic clusters in complex networks. *World Wide Web* 18, 4, 1093–1122. DOI: https://doi.org/10.1007/s11280-014-0310-4

[74] Akshi Kumar and Abha Jain. 2010. *An Algorithmic Framework for Collaborative Interest Group Construction*. Springer, Berlin, 500–508. DOI: https://doi.org/10.1007/978-3-642-14493-6_51

[75] Varun Kumar and Niranjan Pedanekar. 2016. Mining shapes of expertise in online social Q&A communities. In *Proceedings of the 19th ACM Conference on Computer Supported Cooperative Work and Social Computing Companion (CSCW'16 Companion)*. ACM, New York, NY, 317–320. DOI: https://doi.org/10.1145/2818052.2869096

[76] Carl Lagoze and Herbert Van de Sompel. 2001. The open archives initiative: Building a low-barrier interoperability framework. In *Proceedings of the 1st ACM/IEEE-CS Joint Conference on Digital Libraries (JCDL'01)*. ACM, New York, NY, 54–62. DOI: https://doi.org/10.1145/379437.379449

[77] A. Latif, M. T. Afzal, and K. Tochtermann. 2010. Constructing experts profiles from linked open data. In *6th International Conference on Emerging Technologies (ICET'10)*. 33–38. DOI:https://doi.org/10.1109/ICET.2010.5638386

[78] Michael Ley. 2009. DBLP: Some lessons learned. *Proc. VLDB Endow.* 2, 2, 1493–1500. DOI:https://doi.org/10.14778/1687553.1687577

[79] Chunshan Li, William K. Cheung, Yunming Ye, Xiaofeng Zhang, Dianhui Chu, and Xin Li. 2015. The author-topic-community model for author interest profiling and community discovery. *Knowledge and Information Systems* 44, 2, 359–383. DOI:https://doi.org/10.1007/s10115-014-0764-9

[80] Hua Li, Daniel J. T. Powell, Mark Clark, Tifani O'Brien, and Rafael Alonso. 2015. User modeling of skills and expertise from resumes. In *Proceedings of the International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management (IC3K'15)*. SCITEPRESS - Science and Technology Publications, Lda, Portugal, 229–233. DOI:https://doi.org/10.5220/0005622202290233

[81] Juanzi Li, Jie Tang, Jing Zhang, Qiong Luo, Yunhao Liu, and Mingcai Hong. 2008. Arnetminer: Expertise oriented search using social networks. *Frontiers of Computer Science in China* 2, 1, 94–105. DOI:https://doi.org/10.1007/s11704-008-0008-9

[82] Jing Li, Feng Xia, Wei Wang, Zhen Chen, Nana Yaw Asabere, and Huizhen Jiang. 2014. ACRec: A co-authorship based random walk model for academic collaboration recommendation. In *Proceedings of the 23rd International Conference on World Wide Web (WWW'14 Companion)*. ACM, New York, NY, 1209–1214. DOI:https://doi.org/10.1145/2567948.2579034

[83] Wen Li, Carsten Eickhoff, and Arjen P. de Vries. 2016. Probabilistic local expert retrieval. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*. Vol. 9626. 227–239. DOI:https://doi.org/10.1007/978-3-319-30671-1_17 arxiv:1601.02376

[84] Y. Li, S. Ma, Y. Zhang, and R. Huang. 2012. Expertise network discovery via topic and link analysis in online communities. In *IEEE 12th International Conference on Advanced Learning Technologies (ICALT'12)*. 311–315. DOI:https://doi.org/10.1109/ICALT.2012.80

[85] Y. Li and J. Tang. 2008. Expertise search in a time-varying social network. In *9th International Conference on Web-Age Information Management (WAIM'08)*. 293–300. DOI:https://doi.org/10.1109/WAIM.2008.100

[86] Shangsong Liang and Maarten de Rijke. 2016. Formal language models for finding groups of experts. *Information Processing and Management* 52, 4, 529–549. DOI:https://doi.org/10.1016/j.ipm.2015.11.005

[87] Ruud Liebregts and Toine Bogers. 2009. *Design and Evaluation of a University-Wide Expert Search Engine*. Springer, Berlin, 587–594. DOI:https://doi.org/10.1007/978-3-642-00958-7_54

[88] Lili Lin, Zhuoming Xu, Ying Ding, and Xiaozhong Liu. 2013. Finding topic-level experts in scholarly networks. *Scientometrics* 97, 3, 797–819. DOI:https://doi.org/10.1007/s11192-013-0988-6

[89] Shuyi Lin, Wenxing Hong, Dingding Wang, and Tao Li. 2017. A survey on expert finding techniques. *Journal of Intelligent Information Systems* 49, 2 (2017), 1–25. DOI:https://doi.org/10.1007/s10844-016-0440-5

[90] D. Liu, L. Wang, J. Zheng, K. Ning, and L. J. Zhang. 2013. Influence analysis based expert finding model and its applications in enterprise social network. In *IEEE International Conference on Services Computing*. IEEE, 368–375. DOI:https://doi.org/10.1109/SCC.2013.72

[91] Duen-Ren Liu, Yu-Hsuan Chen, Wei-Chen Kao, and Hsiu-Wen Wang. 2013. Integrating expert profile, reputation and link analysis for expert finding in question-answering websites. *Information Processing & Management* 49, 1, 312–329. DOI:https://doi.org/10.1016/j.ipm.2012.07.002

[92] Jingyuan Liu, Debing Liu, Xingyu Yan, Li Dong, Ting Zeng, Yutao Zhang, and Jie Tang. 2014. AMiner-mini: A people search engine for university. In *Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management (CIKM'14)*. ACM, New York, NY, 2069–2071. DOI:https://doi.org/10.1145/2661829.2661852

[93] Ping Liu, Jayne Curson, and Peter Dew. 2002. *Exploring RDF for Expertise Matching within an Organizational Memory*. Springer, Berlin, 100–116. DOI:https://doi.org/10.1007/3-540-47961-9_10

[94] P. Liu, K. Liu, and J. Liu. 2007. Ontology-based expertise matching system within academia. In *Proceedings of the International Conference on Wireless Communications, Networking and Mobile Computing*. IEEE, 5431–5434. DOI:https://doi.org/10.1109/WICOM.2007.1330

[95] Xiaomo Liu, G. Alan Wang, Aditya Johri, Mi Zhou, and Weiguo Fan. 2014. Harnessing global expertise: A comparative study of expertise profiling methods for online communities. *Information Systems Frontiers* 16, 4, 715–727. DOI:https://doi.org/10.1007/s10796-012-9385-6

[96] Zhu Liu, Kan Li, and Dacheng Qu. 2017. Knowledge graph based question routing for community question answering. *Neural Information Processing* 10638, 721–730. DOI:https://doi.org/10.1007/978-3-319-70139-4_73

[97] Ngoc Tu Luong, Tuong Tri Nguyen, Jason J. Jung, and Dosam Hwang. 2015. Discovering co-author relationship in bibliographic data using similarity measures and random walk model. In *Intelligent Information and Database Systems*, Ngoc Thanh Nguyen, Bogdan Trawiński, and Raymond Kosala (Eds.). Springer International Publishing, Cham, 127–136.

[98] Craig Macdonald and Iadh Ounis. 2009. Searching for expertise: Experiments with the voting model. *Comput. J.* 52, 7, 729–748. DOI: https://doi.org/10.1093/comjnl/bxm112

[99] Vitor Mangaravite and Rodrygo L. T. Santos. 2016. On information-theoretic document-person associations for expert search in academia. In *Proceedings of the 39th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'16)*. ACM, New York, NY, 925–928. DOI: https://doi.org/10.1145/2911451.2914751

[100] Vitor Mangaravite, Rodrygo L. T. Santos, Isac S. Ribeiro, Marcos André Gonçalves, and Alberto H. F. Laender. 2016. The LExR collection for expertise retrieval in academia. In *Proceedings of the 39th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'16)*. ACM, New York, NY, 721–724. DOI: https://doi.org/10.1145/2911451.2914678

[101] Christopher D. Manning, Prabhakar Raghavan, and Hinrich Schütze. 2008. *Introduction to Information Retrieval*. Cambridge University Press, New York, NY.

[102] David Mimno and Andrew McCallum. 2007. Expertise modeling for matching papers with reviewers. In *Proceedings of the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD'07)*. ACM, New York, NY, 500–509. DOI: https://doi.org/10.1145/1281192.1281247

[103] Haikun Mou, Qian Geng, Jian Jin, and Chong Chen. 2015. An author subject topic model for expert recommendation. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*. Vol. 9460. 83–95. DOI: https://doi.org/10.1007/978-3-319-28940-3_7

[104] Nasir Naveed, Sergej Sizov, and Steffen Staab. 2011. *ATTention: Understanding Authors and Topics in Context of Temporal Evolution*. Springer, Berlin, 733–737. DOI: https://doi.org/10.1007/978-3-642-20161-5_82

[105] Mahmood Neshati, Hamid Beigy, and Djoerd Hiemstra. 2014. Expert group formation using facility location analysis. *Information Processing and Management* 50, 2, 361–383. DOI: https://doi.org/10.1016/j.ipm.2013.10.001

[106] Mahmood Neshati, Zohreh Fallahnejad, and Hamid Beigy. 2017. On dynamicity of expert finding in community question answering. *Information Processing & Management* 53, 5, 1026–1042. DOI: https://doi.org/10.1016/j.ipm.2017.04.002

[107] Mahmood Neshati, Seyyed Hadi Hashemi, and Hamid Beigy. 2014. Expertise finding in bibliographic network: Topic dominance learning approach. *IEEE Transactions on Cybernetics* 44, 12, 2646–2657. DOI: https://doi.org/10.1109/TCYB.2014.2312614

[108] Mahmood Neshati, Djoerd Hiemstra, Ehsaneddin Asgari, and Hamid Beigy. 2014. Integration of scientific and social networks. *World Wide Web* 17, 5, 1051–1079. DOI: https://doi.org/10.1007/s11280-013-0229-1

[109] Alexander G. Ororbia, II, Jian Wu, Madian Khabsa, Kyle Williams, and Clyde Lee Giles. 2015. Big scholarly data in CiteSeerX: Information extraction from the web. In *Proceedings of the 24th International Conference on World Wide Web (WWW'15 Companion)*. ACM, New York, NY, 597–602. DOI: https://doi.org/10.1145/2740908.2741736

[110] Francesco Osborne, Enrico Motta, and Paul Mulholland. 2013. *Exploring Scholarly Data with Rexplore*. Springer, Berlin, 460–477. DOI: https://doi.org/10.1007/978-3-642-41335-3_29

[111] Aditya Pal. 2015. Discovering experts across multiple domains. In *Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'15)*. ACM, New York, NY, 923–926. DOI: https://doi.org/10.1145/2766462.2767774

[112] Gustavo A. Parada, Hector G. Ceballos, Francisco J. Cantu, and Lucia Rodriguez-Aceves. 2013. Recommending intra-institutional scientific collaboration through coauthorship network visualization. In *Proceedings of the Workshop on Computational Scientometrics: Theory and Applications (CompSci'13)*. ACM, New York, NY, 7–12. DOI: https://doi.org/10.1145/2508497.2508499

[113] T. Peng, D. Zhang, X. Liu, S. Wang, and W. Zuo. 2013. Central author mining from co-authorship network. In *6th International Symposium on Computational Intelligence and Design*, Vol. 1. 228–232. DOI: https://doi.org/10.1109/ISCID.2013.64

[114] Ravikarn Punnarut and Gridaphat Sriharee. 2010. A researcher expertise search system using ontology-based data mining. In *Proceedings of the 7th Asia-Pacific Conference on Conceptual Modelling - Volume 110 (APCCM'10)*. Australian Computer Society, Inc., Darlinghurst, Australia, 71–78. http://dl.acm.org/citation.cfm?id=1862330.1862341.

[115] Isac S. Ribeiro, Rodrygo L. T. Santos, Marcos A. Gonçalves, and Alberto H. F. Laender. 2015. On tag recommendation for expertise profiling: A case study in the scientific domain. In *Proceedings of the 8th ACM International Conference on Web Search and Data Mining (WSDM'15)*. ACM, New York, NY, 189–198. DOI: https://doi.org/10.1145/2684822.2685320

[116] Jan Rybak, Krisztian Balog, and Kjetil Nørvåg. 2014. *Temporal Expertise Profiling*. Springer International Publishing, Cham, 540–546. DOI: https://doi.org/10.1007/978-3-319-06028-6_54

[117] Pavel Serdyukov, Henning Rode, and Djoerd Hiemstra. 2008. Modeling multi-step relevance propagation for expert finding. In *Proceedings of the 17th ACM Conference on Information and Knowledge Management (CIKM'08)*. ACM, New York, NY, 1133–1142. DOI: https://doi.org/10.1145/1458082.1458232

[118] Pavel Serdyukov, Mike Taylor, Vishwa Vinay, Matthew Richardson, and Ryen W. White. 2011. Automatic people tagging for expertise profiling in the enterprise. In *Proceedings of the 33rd European Conference on Advances in Information Retrieval (ECIR'11)*. Springer, Berlin, 399–410. http://dl.acm.org/citation.cfm?id=1996889.1996941.

[119] Chuan Shi, Xiangnan Kong, Philip S. Yu, Sihong Xie, and Bin Wu. 2012. Relevance search in heterogeneous networks. In *Proceedings of the EDBT - International Conference on Extending Database Technology*. ACM, 180–191. DOI: https://doi.org/10.1145/2247596.2247618

[120] Jasmeet Singh and Vishal Gupta. 2016. Text stemming: Approaches, applications, and challenges. *ACM Comput. Surv.* 49, 3, Article 45, 46 pages. DOI: https://doi.org/10.1145/2975608

[121] Elena Smirnova and Krisztian Balog. 2011. A user-oriented model for expert finding. In *Proceedings of the 33rd European Conference on Advances in Information Retrieval (ECIR'11)*. Springer, Berlin, 580–592. http://dl.acm.org/citation.cfm?id=1996889.1996964.

[122] Mark Steyvers, Padhraic Smyth, Michal Rosen-Zvi, and Thomas Griffiths. 2004. Probabilistic author-topic models for information discovery. In *Proceedings of the 10th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD'04)*. ACM, New York, NY, 306–315. DOI: https://doi.org/10.1145/1014052.1014087

[123] Yizhou Sun, Rick Barber, Manish Gupta, Charu C. Aggarwal, and Jiawei Han. 2011. Co-author relationship prediction in heterogeneous bibliographic networks. In *Proceedings of the International Conference on Advances in Social Networks Analysis and Mining (ASONAM'11)*. IEEE Computer Society, Washington, DC, 121–128. DOI: https://doi.org/10.1109/ASONAM.2011.112

[124] Jie Tang. 2016. AMiner: Toward understanding big scholar data. In *Proceedings of the 9th ACM International Conference on Web Search and Data Mining (WSDM'16)*. ACM, New York, NY, 467–467. DOI: https://doi.org/10.1145/2835776.2835849

[125] J. Tang, R. Jin, and J. Zhang. 2008. A topic modeling approach and its integration into the random walk framework for academic search. In *8th IEEE International Conference on Data Mining*. IEEE,1055–1060. DOI: https://doi.org/10.1109/ICDM.2008.71

[126] Jie Tang, Sen Wu, Jimeng Sun, and Hang Su. 2012. Cross-domain collaboration recommendation. In *Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD'12)*. ACM, New York, NY, 1285–1293. DOI: https://doi.org/10.1145/2339530.2339730

[127] Jie Tang, Jing Zhang, Ruoming Jin, Zi Yang, Keke Cai, Li Zhang, and Zhong Su. 2011. Topic level expertise search over heterogeneous networks. *Machine Learning* 82, 2, 211–237. DOI: https://doi.org/10.1007/s10994-010-5212-9

[128] Jie Tang, Jing Zhang, Limin Yao, Juanzi Li, Li Zhang, and Zhong Su. 2008. ArnetMiner: Extraction and mining of academic social networks. In *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD'08)*. ACM, New York, NY, 990–998. DOI: https://doi.org/10.1145/1401890.1402008

[129] Wenbin Tang, Jie Tang, and Chenhao Tan. 2010. Expertise matching via constraint-based optimization. In *Proceedings of the IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology - Volume 01 (WI-IAT'10)*. IEEE Computer Society, Washington, DC, 34–41. DOI: https://doi.org/10.1109/WI-IAT.2010.133

[130] Quan Thanh Tho, Siu Cheng Hui, and A. C. M. Fong. 2003. A web mining approach for finding expertise in research areas. In *Proceedings of the International Conference on Cyberworlds*. 310–317. DOI: https://doi.org/10.1109/CYBER.2003.1253470

[131] Yuancheng Tu, Nikhil Johri, Dan Roth, and Julia Hockenmaier. 2010. Citation author topic model in expert search. In *Proceedings of the International Conference on Computational Linguistics*. Association for Computational Linguistics, 1265–1273. http://dl.acm.org/citation.cfm?id=1944566.1944711.

[132] David van Dijk, Manos Tsagkias, and Maarten de Rijke. 2015. Early detection of topical expertise in community question answering. *Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'15)*. ACM, 995–998. DOI: https://doi.org/10.1145/2766462.2767840

[133] Christophe Van Gysel, Maarten de Rijke, and Evangelos Kanoulas. 2017. Structural regularities in text-based entity vector spaces. DOI: https://doi.org/10.1145/3121050.3121066 arxiv:1707.07930

[134] Christophe Van Gysel, Maarten de Rijke, and Marcel Worring. 2016. Unsupervised, efficient and semantic expertise retrieval. In *Proceedings of the WWW - International Conference on World Wide Web*. International World Wide Web Conferences Steering Committee, 1069–1079. DOI: https://doi.org/10.1145/2872427.2882974

[135] Joris Vertommen, Frizo Janssens, Bart De Moor, and Joost R. Duflou. 2008. Multiple-vector user profiles in support of knowledge sharing. *Inf. Sci.* 178, 17, 3333–3346. DOI: https://doi.org/10.1016/j.ins.2008.05.001

[136] Jianwen Wang, Xiaohua Hu, Xinhui Tu, and Tingting He. 2012. Author-conference topic-connection model for academic network search. In *Proceedings of the 21st ACM International Conference on Information and Knowledge Management (CIKM'12)*. ACM, New York, NY, 2179–2183. DOI: https://doi.org/10.1145/2396761.2398597

[137] Xiaolong Wang, Chengxiang Zhai, and Dan Roth. 2013. Understanding evolution of research themes: A probabilistic generative model for citations. In *Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD'13)*. ACM, New York, NY, 1115–1123. DOI: https://doi.org/10.1145/2487575.2487698

[138] Xiaoqin Xie, Yijia Li, Zhiqiang Zhang, Haiwei Pan, and Shuai Han. 2016. *A Topic-Specific Contextual Expert Finding Method in Social Network*. Springer International Publishing, Cham, 292–303. DOI: https://doi.org/10.1007/978-3-319-45814-4_24

[139] Shuo Xu, Qingwei Shi, Xiaodong Qiao, Lijun Zhu, Hanmin Jung, Seungwoo Lee, and Sung-Pil Choi. 2014. Author-topic over time (AToT): A dynamic users' interest model. In *Mobile, Ubiquitous, and Intelligent Computing: MUSIC 2013*, James J. (Jong Hyuk) Park, Hojjat Adeli, Namje Park, and Isaac Woungang (Eds.). Springer, Berlin, 239–245. DOI: https://doi.org/10.1007/978-3-642-40675-1_37

[140] Yunhong Xu, Xitong Guo, Jinxing Hao, Jian Ma, Raymond Y. K. Lau, and Wei Xu. 2012. Combining social network and semantic concept analysis for personalized academic researcher recommendation. *Decision Support Systems* 54, 1, 564–573. DOI: https://doi.org/10.1016/j.dss.2012.08.003

[141] Chen Yang, Jian Ma, Thushari Silva, Xiaoyan Liu, and Zhongsheng Hua. 2014. A multilevel information mining approach for expert recommendation in online scientific communities. *Computer Journal* 58, 9, 1921–1936. DOI: https://doi.org/10.1093/comjnl/bxu033

[142] C. Yang, J. Sun, J. Ma, S. Zhang, G. Wang, and Z. Hua. 2015. Scientific collaborator recommendation in heterogeneous bibliographic networks. In *48th Hawaii International Conference on System Sciences*. 552–561. DOI: https://doi.org/10.1109/HICSS.2015.73

[143] Kun-Woo Yang and Soon-Young Huh. 2008. Automatic expert identification using a text categorization technique in knowledge management systems. *Expert Systems with Applications* 34, 2, 1445–1455. DOI: https://doi.org/10.1016/j.eswa.2007.01.010

[144] Zaihan Yang, Liangjie Hong, and Brian D. Davison. 2013. Academic network analysis: A joint topic modeling approach. In *Proceedings of the IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM'13)*. ACM, New York, NY, 324–333. DOI: https://doi.org/10.1145/2492517.2492524

[145] S. Zehnalova, Z. Horak, M. Kudelka, and V. Snasel. 2012. Evolution of author's topic in authorship network. In *IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*. ACM, New York, NY, 1207–1210. DOI: https://doi.org/10.1109/ASONAM.2012.208

[146] Zhenjiang Zhan, Lichun Yang, Shenghua Bao, Dingyi Han, Zhong Su, and Yong Yu. 2011. Finding appropriate experts for collaboration. In *Proceedings of the 12th International Conference on Web-age Information Management (WAIM'11)*. Springer, Berlin,327–339. http://dl.acm.org/citation.cfm?id=2035562.2035601.

[147] Zhou Zhao, Qifan Yang, Deng Cai, Xiaofei He, and Yueting Zhuang. 2016. Expert finding for community-based question answering via ranking metric network learning. In *Proceedings of the 25th International Joint Conference on Artificial Intelligence (IJCAI'16)*. AAAI Press, 3000–3006. http://dl.acm.org/citation.cfm?id=3060832.3061041.

[148] Guoqing Zheng, Jinwen Guo, Lichun Yang, Shengliang Xu, Shenghua Bao, Zhong Su, Dingyi Han, and Yong Yu. 2011. Mining topics on participations for community discovery. In *Proceedings of the ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM, 445–454. DOI: https://doi.org/10.1145/2009916.2009977

[149] Xing Zhou, Lixin Ding, Zhaokui Li, and Runze Wan. 2017. Collaborator recommendation in heterogeneous bibliographic networks using random walks. *Information Retrieval Journal* 20, 4 (2017), 1–21. DOI: https://doi.org/10.1007/s10791-017-9300-3

[150] Hengshu Zhu, Enhong Chen, Hui Xiong, Huanhuan Cao, and Jilei Tian. 2014. Ranking user authority with relevant knowledge categories for expert finding. *World Wide Web* 17, 5, 1081–1107. DOI: https://doi.org/10.1007/s11280-013-0217-5

[151] J. Zhu, A. L. Goncalves, V. S. Uren, E. Motta, and R. Pacheco. 2005. Mining Web data for competency management. In *IEEE/WIC/ACM International Conference on Web Intelligence (WI'05)*. 94–100. DOI: https://doi.org/10.1109/WI.2005.99

[152] Jianhan Zhu, Xiangji Huang, Dawei Song, and Stefan Rüger. 2010. Integrating multiple document features in language models for expert finding. *Knowl. Inf. Syst.* 23, 1, 29–54. DOI: https://doi.org/10.1007/s10115-009-0202-6

Received February 2018; revised March 2019; accepted May 2019