# Experion: A framework for contextualizing evidence in expert finding

**Rodrigo Gonçalves**[1], **Carina F. Dorneles**[1]

[1]Departamento de Informática e Estatística
Universidade Federal de Santa Catarina (UFSC)
Caixa Postal 476 – 88.040-900 – Florianópolis – SC – Brazil

{rodrigo.g,carina.dorneles}@ufsc.br

***Abstract.*** *Expert finding is traditionally related to a subject of research in information retrieval and, often, is taken to mean "expertise retrieval within a specific organization" [Balog et al. 2012]. The task involves finding an expert in an expertise topic. Even though there are interesting proposals in the literature, they do not consider the context in which a given expertise is bound. This Ph.D. thesis introduces the concept of a framework that chronologically contextualizes search results in expert finding. Our motivation is to provide more accurate results of search processes related to finding experts in a given topic, contextualizing the expertise on professional/academic activities, an open research topic. In this paper, we present the main concepts of the framework we are developing and a general overview of its operation. At the moment, we are using the Lattes platform as a data source, for which we developed a process to extract expertise evidence, supported by the Crossref database.*

## 1. Informações gerais

- **Tese de doutorado:** Experion: A framework for contextualizing evidence in expert finding.
- **Estudante:** Rodrigo Gonçalves.
- **Orientadora:** Dra. Carina Friedrich Dorneles.
- **Mês e ano de ingresso:** Março/2016 (com trancamentos nos semestres 2017/1 e 2018/2).
- **Mês e ano previsto para defesa:** Agosto/2022.
- **Etapas concluídas:**
  - **Disciplinas Obrigatórias** - Junho/2019
  - **Seminário de Andamento** - Junho/2018
  - **Exame de Qualificação** - Novembro/2019
- **Publicações:** Automated Expertise Retrieval: A Taxonomy-Based Survey and Open Issues [Gonçalves and Dorneles 2019], ACM Computing Surveys.

## 2. Introduction

According to Balog et al. [Balog et al. 2012], *expertise* is a loosely-defined concept, usually referred to as *"tacit knowledge"*, i.e, the knowledge that people acquire through experiences in their lives that is stored in their minds. It is difficult to express such knowledge in a detailed, formalized and complete way that allows other people to know about it. This is a valuable and challenging research topic in data retrieval that deals with finding ways to discover and automatically describe this type of knowledge.

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 [Balog et al. 2012]: authored documents (articles, reports), electronic communications, social networks, etc. The process of finding and extracting this kind of evidence, and linking it to a particular expertise, is called *expertise retrieval*, which has two basic applications: expert finding and expert profiling [Balog et al. 2012]. *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 expertise.

The problem of helping the user understand the results of an expertise retrieval process is often ignored in the literature. Current approaches usually give a list of people or a graph cluster as a result but fail to describe how they obtained this data [Chen et al. 2013, Pal 2015]. By allowing the user to understand the result, he is free to use his judgment and decide whether or not to go ahead with contacting the referred expert. 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 [Gonçalves and Dorneles 2019].

The motivation for our work is to address the gap in current works where they do not provide enough information for the user to understand the context where the expert acquired or applied the expertise of interest. We extend existing expert finding systems where, based on collected data from the expertise evidence, we elaborate the context where the expertise evidences occurred. This provides the user with more information to understand the expertise while reusing/improving existing work in expert finding.

Our framework can be generalized and further applied to any search system in which results can provide some idea of context information, from which one or more contexts can be derived. Thus it can be applied to any search process result in database systems that contains an associated context.

## 3. Related work

In this section we provide a brief overview of existing work in expert finding. For an extensive list of related work see the survey [Gonçalves and Dorneles 2019], which details existing work approaches and introduces, among other open issues, those which were cited in this work as a justification for our research.

Current studies in expert finding adopt several approaches to find experts. Many use document-centric methods, such as: (i) using the SVM to represent and search for experts, given keywords of interest [Chen et al. 2013]; (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 which models that can best generate the query [Pal 2015]; (iii) representing expertise through ontologies and using them to search for experts[Punnarut and Sriharee 2010]; and (iv) clustering documents based on their keywords, allowing the retrieval of experts associated to documents in the same clusters with related keywords [Tho et al. 2003]. However, as already mentioned, the gap in current works is that they do not provide enough information for the user to understand the context where the expert acquired or applied the expertise of interest.

To the best of our knowledge, ScholarLens[Sateli et al. 2017] is the closest related work to our proposal, although it focus on expert profiling instead of expert finding. ScholarLens is a platform that, given a set of articles from a researcher, identifies and extracted named entities (using NLP methods) and, using DBPedia as a support, elaborate a knowledge-database representing the researcher knowledge. The profiles are built using RDF and the competences (expertise) are modeled using the IntelLEO ontology.

## 4. Proposal

Our framework, named **Experion**, proposes contextualizing the result of expert finding systems. In such systems, a list of expertise evidence is located and used to justify a person as an expert in a given topic from a search criterion. The expertise evidence may have contextual information associated with them. The Experion's purpose is to extract such information and elaborate one or more chronological contexts. The main Experion's components are introduced in Figure 1.

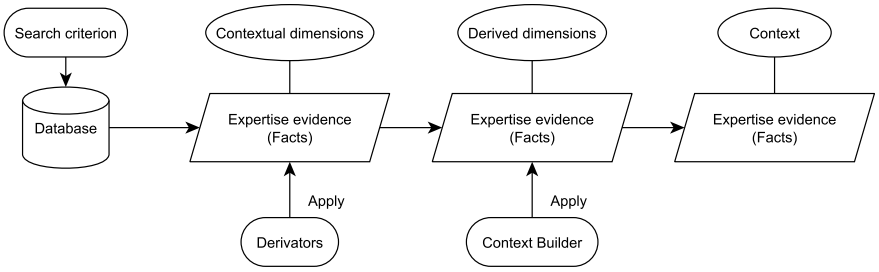


**Figura 1. Experion overview**

Experion starts with a set of expertise evidence associated with one or more Experts. To allow generalization, the framework uses the term *Fact* to refer to expertise evidence and the term *Entity* to refer to the Expert. We assume that the expertise evidence (Facts) contain data that give insights into the context where they occurred. We denominate such data as *contextual dimensions*, and they are used as input for functions that extract qualified information about the context. These functions are called *derivators*. The qualified information obtained by the derivators is stored as *derived dimensions* associated with the expertise evidence (Facts). Lastly, another function, denominated *context builder*, is applied to elaborate the context associated to each Fact. In the following, each of these elements is described.

### 4.1. Entities and Facts

An **Entity** is the final object of interest for the search process. In order to locate these entities, the framework uses intermediary objects associated with these Entities, which

we denominate as **Facts**. Thus the search process starts by locating a set of *Facts* given a search criterion and from these Facts identifies the associated entities.

A **Fact** is a multidimensional object. Each dimension in a fact can be used for searching or for contextualizing. A dimension used for searching is called an *index dimension* and a dimension used for contextualizing is called a *contextual dimension*.

In the context of expert finding, the entities are candidate experts. We can use keywords as a search criterion (related to the expertise of interest) and locate evidence of expertise (Facts) to locate such experts. To locate evidence, we can index them based on their keywords through an inverted index (an index dimension), for example, and use the search criterion (keyword) to locate them. Each Fact (expertise evidence) can have one or more contextual dimensions, such as the institution name in the case of teaching expertise evidence; or the name of an associated for expertise evidence related to a research project.

### 4.2. Context, Derived Dimensions and Derivators

A **Fact** associated with an entity can be part of one or more contexts. A **Context** is an abstraction that situates a given Fact under perspectives of interest in a given period. In expert finding, we can consider as such perspectives: Education Level (Undergraduate, Graduate, Ph.D., etc.); Activity (Education, Teaching, Research, etc.); Cooperation (Individual work, Collective Work); Scope (National, International); Environment (Academic, Enterprise, etc); etc.

To identify such perspectives based on the contextual dimensions we define the concept of **Derivator Function**, or **Derivator** for short. A **Derivator** receives a set of Facts and, based on their contextualized dimensions is capable of producing **Derived Dimensions**, which is the perspective of interest for **Contexts**.

In expert finding, a **Derived Property** could be, for example, “activity”, which can be “teaching” or “research”. Another example could be “degree”, which could be “undergraduate” or “graduate”. An example of **Derivator** could be a function that, given a piece of expertise evidence associated with teaching (a class lecture, for example) and is also associated with a university, can derive an activity of “teaching” and a “degree” of “graduate”. Thus from the two Derived Dimensions (“teaching” and “graduate”) a Context of “graduate teaching” can be established, which describes a **Context** under two perspectives of interest: Education Level and Activity.

### 4.3. Context Builder

As introduced, a **Context** is built from Derived Dimensions obtained from Facts. To formalize the construction of a context, we introduced a function type called **Context Builder**.

A Context Builder receives a set of Facts with their Derived Dimensions (through the application of Derivators) and can produce one or more Contexts, associating each Fact with one or more Contexts at a given degree of confidence (*Confidence factor*). This indicates how well a given Fact corroborates the Context to which it is associated.

## 5. Experion

Figure 2 details the execution of the Experion framework. A contextualized search process starts by executing a query over a database (given a search criterion **(1)**), which

returns a set of Facts **(2)**. A Fact, as defined earlier, is any object that contains index dimensions (which allows the search process to find them) and contextual dimensions, which are used by the framework to build the Contexts. Each Fact is also associated with the Entity of interest for the search process. In this case, a Person or, more specifically, a Candidate Expert. Such association allows separating the facts into groups, as seen in the Figure. Once we have a set of Facts, one or more Derivators are applied over the set **(3)**. As a result of their application, Derived Dimensions are introduced in the Fact objects **(4)**. As described earlier, these dimensions are generated using the information contained in the Fact as well as analyzing the information from other facts from the set, allowing a certain level of inference.

The next step of the framework is to apply a Context Builder over the set of Facts **(5)**. The context builder analyzes the derived dimensions values and establishes one or more Contexts **(6)** associated to one or more Facts. For each pair (Fact,Context), it calculates a Confidence Factor (Cf) **(7)**, which indicates how strongly the given Fact contributes for the Context definition.

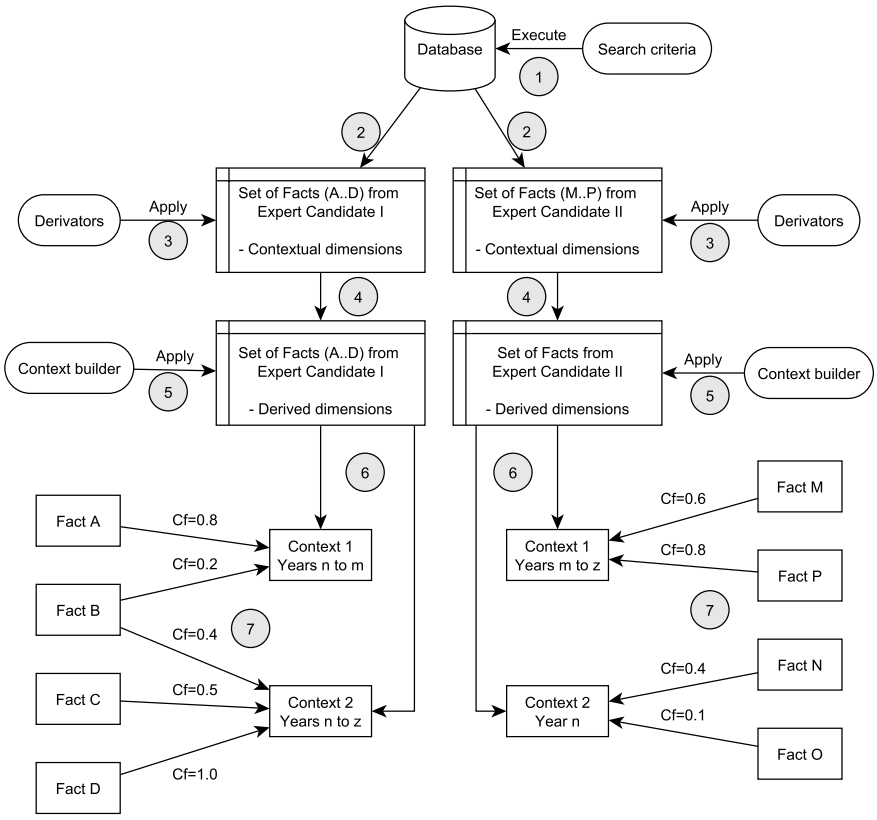


**Figura 2. Contextualized search example**

## 5.1. Expert finding example

To demonstrate the application of Experion, we exemplify how it can be used to contextualize expert finding results, in an academic environment. As introduced earlier, an Entity would be a Candidate Expert and a Fact would be an Expertise Evidence. As examples of Fact and their index/contextual dimensions we cite:

1. **Published Periodic Article:** keywords (index dimension); year, co-authors, periodic, JCR (contextual dimensions).
2. **Academic Advising:** keywords (index dimension); year, student, institution, degree (contextual dimensions).
3. **Research Project:** keywords (index dimension); year, colleague, institution (contextual dimensions).

There are several Derived Properties that can be obtained from expertise evidence. Previously, we have cited five of them. Now, we describe how Derivator Functions could generate them:

1. **Level:** it can not be obtained from a Periodic Article or a Research project. It can be obtained from an Academic Advising (through the level contextual dimension).
2. **Activity:** a Published Periodic Article indicates Research and, if a co-author appears in an Academic Advising evidence as well, indicates Advising; an Academic Advising indicates Orientation per itself; the same happens for Research given a Research Project evidence.
3. **Cooperation:** in the case of Published Periodic Article, if there are more than one co-author this indicates Group cooperation; an Academic Advising can not indicate a cooperation level; a research group per itself is already an indication of a Group cooperation;
4. **Scope:** in the case of a Periodic Article the language used by the keywords can indicate if it is a national/international scope; for Advising it can't indicate; for research project the language used by the keywords can indicate the scope.
5. **Environment:** Periodic Article and Academic Advising indicates an Academic environment; a Research Project can indicate an Academic and/or Enterprise environment based on keywords and related institutions.

As example of a *Context Builder* for Expert Finding, we introduce the *Simple Weighted Context Builder* - SWCB. The SWCB concatenates the derived properties values to build the Contexts and assigns a confidence degree to each Fact associated with a context.

To set up the Contexts, the Context Builder builds a timeline where each interval is the year(s) to which the Facts are associated. Each Fact set of derived properties represents a candidate context for its associated years. If a Fact occurs in a single year its confidence is calculated as 1 for its associated derived properties. If it occurs in more than one year, the confidence factor is divided by the number of years. For each year the Context Builder identifies the sets of derived properties which share properties and clusters them by similarity. The result of the clustering process is the context associated to the Facts in that year.

## 6. Contributions and Future Work

In this work we introduced Experion, a framework proposal aiming at contextualizing expertise evidence in expert finding systems. To validate the proposal, we are developing a reference implementation based on the Lattes curricula [1], complemented with data from the Crossref database [2]. This implementation focus on answering two questions:

[1] `https://lattes.cnpq.br`
[2] `https://www.crossref.org/`

*(i) how can the context be used to improve the expertise analysis:* the Derived Dimensions can be used to weight whether the associated Fact is a good indication of expertise. For example, a PhD Examination Board is a stronger evidence than an Undergraduate. For an Article, the venue could be used as a weight to expertise qualification as well. Compared to existing work, our differs by allowing a dynamic weighting to the expertise evidence, based on the goal of the expertise search process.

*(ii) how to to use the context as a ranking tool:* since the focus on our preliminary tests was on extracting the context associated to the Facts, we developed a basic ranking system which takes into account the kind of Fact found as expertise evidence and the Derived Dimensions associated to calculate a basic ranking value. Our work improves existing work by providing a reasoning behind the ranking process.

Our next steps are divided in four phases:

- Complete the framework implementation and formal definition. Execute preliminary tests to verify that the contexts generated match an human interpretation of the expertise evidence. Out timeline for this phase is 4 months;
- Produce and submit an article to an event with the framework proposal and initial test results - we estimate 2 months;
- Thesis writing, including additional functions in the framework - we reserved 4 months for this phase;
- Defend the thesis and produce and submit an article to a periodic with the thesis result - 2 months are expected to be necessary for this phase.

## Referências


Balog, K., Fang, Y., de Rijke, M., Serdyukov, P., and Si, L. (2012). Expertise retrieval. Foundations and Trends® in Information Retrieval, 6(2–3):127–256.

Chen, H.-H., Treeratpituk, P., Mitra, P., and Giles, C. L. (2013). Csseer: An expert recommendation system based on citeseerx. In Proceedings of the 13th ACM/IEEE-CS Joint Conference on Digital Libraries, JCDL '13, pages 381–382, New York, NY, USA. ACM.

Gonçalves, R. and Dorneles, C. F. (2019). Automated expertise retrieval: A taxonomy-based survey and open issues. ACM Comput. Surv., 52(5).

Pal, A. (2015). Discovering experts across multiple domains. In Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '15, pages 923–926, New York, NY, USA. ACM.

Punnarut, R. and Sriharee, G. (2010). A researcher expertise search system using ontology-based data mining. In Proceedings of the Seventh Asia-Pacific Conference on Conceptual Modelling - Volume 110, APCCM '10, pages 71–78, Darlinghurst, Australia, Australia. Australian Computer Society, Inc.

Sateli, B., Löffler, F., König-Ries, B., and Witte, R. (2017). Scholarlens: Extracting competences from research publications for the automatic generation of semantic user profiles. PeerJ Computer Science, 2017.

Tho, Q. T., Hui, S. C., and Fong, A. C. M. (2003). A web mining approach for finding expertise in research areas. In Proceedings. 2003 International Conference on Cyberworlds, pages 310–317.