UNIVERSIDADE FEDERAL DE SANTA CATARINA
CAMPUS FLORIANÓPOLIS
PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIA DA COMPUTAÇÃO

Rodrigo Gonçalves

Experion - a framework for expertise retrieval: fact-based context injection in expert finding systems for multiple contextualized result interpretations

Florianópolis
2023

Rodrigo Gonçalves

Experion - a framework for expertise retrieval: fact-based context injection in expert finding systems for multiple contextualized result interpretations

Tese submetida ao Programa de Pós-Graduação em Ciência da Computação da Universidade Federal de Santa Catarina para a obtenção do título de doutor em Ciência da Computação.
Orientadora: Prof.(a) Carina Friedrich Dorneles, Dr.(a)

Florianópolis
2023

Ficha de identificação da obra elaborada pelo autor,
através do Programa de Geração Automática da Biblioteca Universitária da UFSC.

Gonçalves, Rodrigo
Experion - a framework for expertise retrieval : fact based context injection in expert finding systems for multiple contextualized result interpretations / Rodrigo Gonçalves ; orientadora, Carina Friedrich Dorneles, 2023.
117 p.

Tese (doutorado) - Universidade Federal de Santa Catarina, Centro Tecnológico, Programa de Pós-Graduação em Ciência da Computação, Florianópolis, 2023.

Inclui referências.

1. Ciência da Computação. 2. recuperação de expertise. 3. descoberta de especialista. 4. contextualização. I. Dorneles, Carina Friedrich. II. Universidade Federal de Santa Catarina. Programa de Pós-Graduação em Ciência da Computação. III. Título.

Rodrigo Gonçalves

Experion - a framework for expertise retrieval: fact-based context injection in expert finding systems for multiple contextualized result interpretations

O presente trabalho em nível de doutorado foi avaliado e aprovado por banca examinadora composta pelos seguintes membros:

Prof.(a) Renata Galante, Dr.(a)
Universidade Federal do Rio Grande do Sul

Prof. Rodrygo Santos, Dr.
Universidade Federal de Minas Gerais

Prof. Ronaldo dos Santos Mello, Dr.
Universidade Fededral de Santa Catarina

Certificamos que esta é a **versão original e final** do trabalho de conclusão que foi julgado adequado para obtenção do título de doutor em Ciência da Computação.

Documento assinado digitalmente
PATRICIA DELLA MEA PLENTZ
Data: 06/02/2023 17:33:29-0300
CPF: ***.243.700-**
Verifique as assinaturas em https://v.ufsc.br

Coordenação do Programa de
Pós-Graduação

Documento assinado digitalmente
CARINA FRIEDRICH DORNELES
Data: 06/02/2023 14:42:17-0300
CPF: ***.834.250-**
Verifique as assinaturas em https://v.ufsc.br

Prof.(a) Carina Friedrich Dorneles, Dr.(a)
Orientadora

Florianópolis, 2023.

This work is dedicated to my father, *in memoriam*, and to my mother, who always encouraged me to study and seek knowledge.

ACKNOWLEDGEMENTS

Thanks to everyone who helped on this journey.

RESUMO

A recuperação de expertise identifica e descreve automaticamente a expertise associada a uma pessoa. A expertise é gerada com base nas evidências (publicações e atividades, por exemplo) associadas à pessoa. Após uma longa revisão de trabalhos existentes, que produziu uma taxonomia facetada e identificou várias questões em aberto, focou-se neste trabalho em melhorar a compreensão do usuário sobre os resultados produzidos pelos sistemas de *descoberta de especialistas*. A descoberta de especialistas lida com, dada uma expertise de interesse, localizar possíveis especialistas na mesma. A hipótese de pesquisa é de que, ao incluir informações contextuais nos resultados destes sistemas, os usuários entenderão melhor os mesmos. Os trabalhos atuais produzem apenas uma lista de possíveis especialistas. A lista não contém contexto ou, no máximo, contém uma contextualização limitada, focada apenas na especialidade em si. Ao buscar um especialista, outros contextos podem desempenhar um papel essencial na escolha da pessoa certa - para uma posição de lecionar uma disciplina universitária, alguém com experiência anterior em ensino é mais desejável do que alguém sem experiência em ensino. Este trabalho apresenta o **Experion**, um framework que padroniza e permite a contextualização de evidências de expertise. Ele identifica, coleta e analisa dados contextuais associados às evidências, como onde, quando e com quem determinada evidência de expertise ocorreu ou foi produzida. Essa análise gera um conjunto padronizado de dados contextuais. O Experion, usando tais dados, descreve automaticamente o contexto para a evidência de expertise. Aplicou-se o Experion aos dados da plataforma Lattes e desenvolveu-se um sistema de busca de especialistas que contextualiza seus resultados usando o framework. Com esse sistema, foram promovidos experimentos qualitativos com usuários e coletados seus feedbacks, que confirmaram a hipótese de que a contextualização melhora a compreensão dos resultados.

Palavras-chave: recuperação de expertise, descoberta de especialista, contextualização.

RESUMO EXPANDIDO

INTRODUÇÃO

Expertise pode ser vagamente definida como o conhecimento que as pessoas adquirem por meio de experiências de vida (BALOG et al., 2012). A recuperação de expertise lida com a descoberta automática e a descrição adequada desse tipo de conhecimento. De acordo com Balog et al. (BALOG et al., 2012), *expertise* é um conceito vagamente definido que não é fácil de formalizar ou representar e é geralmente referido como *"conhecimento tácito,"* ou seja, o conhecimento que as pessoas adquirem através de experiências em suas vidas, que está armazenado em suas mentes. Uma forma de perceber o conhecimento tácito é analisar as *evidências de expertise* associadas a uma pessoa. *Evidência de expertise* é um artefato que contém informações relacionadas à expertise de uma pessoa (BALOG et al., 2012).

Existem muitas fontes de dados de onde esses artefatos podem ser obtidos: documentos de autoria da pessoa (artigos, relatórios), comunicações eletrônicas e redes sociais. Uma vez que é demorado e complexo recuperar e descrever tais conhecimentos, abordagens automatizadas tornaram-se um tópico de pesquisa interessante para muitas comunidades de ciência da computação nos últimos anos. A descoberta de especialistas e a descrição do perfil de expertise são as duas principais aplicações de recuperação de expertise (BALOG et al., 2012). Em *descoberta de especialistas*, dada uma lista de um ou mais tópicos de interesse, localizam-se os especialistas nestes tópicos. *Descrição do perfil de expertise* envolve a construção de perfis de expertise, ou seja, descrições da expertise das pessoas (BALOG; DE RIJKE, 2007).

Os trabalhos existentes na área de localização de especialistas variam em suas técnicas para elaborar e representar a expertise associada a uma pessoa. Ao apresentar seus resultados, os sistemas existentes carecem de uma representação da expertise identificada ou apresentam representações limitadas. Tais representações: (i) não descrevem a evidência de expertise sobre a qual a expertise foi elaborada e; (ii) não explicam como a expertise foi obtida ou demonstrada - o *contexto* associado a uma expertise.

O contexto pode indicar, por exemplo, onde uma pessoa obteve ou demonstrou tal expertise e com quem. Ter um contexto ajuda a evitar que para determinada tarefa seja selecionado um profissional com a expertise necessária em um tema, mas não no contexto desejado - por exemplo, selecionar um pesquisador que nunca lecionou para atuar como professor de um determinado curso. Melhorar os sistemas existentes de descoberta de especialistas sobre a ótica destas limitações é a principal motivação deste trabalho.

OBJETIVOS

Este trabalho melhora a apresentação e compreensão dos resultados em sistemas de descoberta de especialistas. Ele apresenta um novo framework chamado Experion, que fornece uma representação contextualizada de evidências de expertise. Os sistemas de descoberta de especialistas podem usar essa representação para apresentar seus resultados a um usuário.

Estão contemplados os seguintes objetivos específicos: (i) propõe-se uma representação padrão para informações sobre expertise; (ii) gerar uma representação humana-

mente compreensível de informações sobre expertise e incluir informações contextuais em sistemas de descoberta de especialistas por meio da coleta de dados associados às evidências de expertise; (iii) validar o benefício de resultados mais detalhados em sistemas de descoberta de especialistas aplicando o framework na plataforma Lattes e promovendo experimentos e; (iv) melhorar a informação contextual disponível associada às evidências de expertise com um método de auto-ajuste de injeção de contexto nas evidências de expertise.

METODOLOGIA

A metodologia de pesquisa adotada neste trabalho compreendeu sete fases: (i) realização de uma revisão da literatura; (ii) elaborar um survey sobre recuperação de expertise; (iii) estabelecer e desenvolver uma proposta; (iv) implementação de uma ferramenta protótipo para testar e validar a proposta; (v) coleta de feedback de usuários sobre o protótipo desenvolvido; (vi) promover maior otimização da proposta, adequando o protótipo da ferramenta e; (vii) coletar feedback adicional dos usuários e analisar os resultados.
Concluída a revisão da literatura, um *survey* sobre recuperação de expertise foi elaborado e publicado (GONÇALVES; DORNELES, C. F., 2019). Este *survey* teve uma abordagem mais geral e pretendia confirmar a compreensão do estado da arte e incluiu uma nova taxonomia para classificar os trabalhos existentes.
A proposta de trabalho e os objetivos específicos associados foram formulados com o estado da arte compreendido, incluindo questões em aberto. A proposta foi desenvolvida, e um protótipo - um sistema de descoberta de especialistas - foi implementado para validar a proposta e atingir os objetivos específicos associados. Foi promovido um experimento onde foi solicitado a um grupo de usuários a utilização do sistema desenvolvido. Cada usuário respondeu então um questionário sobre o impacto das informações adicionais e contextualização da expertise, fornecidas pela Experion, na compreensão dos resultados do sistema de descoberta de especialistas.
Com base no feedback dos usuários no experimento, uma otimização adicional a proposta original foi desenvolvida, chamada injeção de contexto. Ele usa os dados contextuais disponíveis na evidência de expertise para melhorar o contexto de evidências que carecem de informações contextuais. Esta otimização exigiu maiores desenvolvimentos na ferramenta protótipo, que foi então submetida a um experimento qualitativo. Entrevistou-se três especialistas com mais de dez anos de experiência em suas áreas, solicitando que usassem a ferramenta desenvolvida e coletou-se suas impressões e sugestões. As entrevistas foram semiestruturadas, com roteiro básico que permitia aos especialistas manipular a ferramenta livremente, e duraram em torno de uma hora cada. Auxílio e maiores esclarecimentos foram prestados durante a entrevista sobre a ferramenta e o framework. Com base neste experimento, alcançou-se a proposta de tese e estabeleceu-se trabalhos futuros para melhorar os resultados.

RESULTADOS E DISCUSSÃO

Foram geradas duas novas contribuições para a Recuperação de Expertise neste trabalho: uma taxonomia facetada e o framework Experion. A taxonomia proposta classifica os trabalhos de Recuperação de Expertise em várias perspectivas, como que tipo de fonte de dados é usada, quais técnicas são usadas e qual é a aplicação final

(descoberta de especialistas, descrição de perfil de expertise, entre outros). O Framework Experion permite contextualizar as evidências de expertise para a descoberta de especialistas.
Inicialmente, com base na taxonomia apresentada, foi promovido um extenso levantamento sobre o estado da arte da recuperação de expertise. Com base nesse levantamento, várias questões em aberto foram identificadas e analisadas. Dentre as questões em aberto, focou-se em duas questões: contextualização e explicação dos resultados.
Para fornecer contextualização na descoberta de especialistas, desenvolveu-se um Framework chamado Experion. Este framework compreende um conjunto de entidades (conceitos) - *Entidade*, *Fato*, *Dimensão* e *Contexto* - e funções para construir o contexto de expertise - Funções *Derivadoras* e *Construtores de Contexto*. Foi desenvolvida uma aplicação do Experion no contexto de expertise na Academia, descrevendo em detalhe a implementação do framework nesse contexto.
Usando o framework Experion, qualquer sistema de descoberta de especialistas pode extrair as informações de contexto associadas a um conjunto de evidências de expertise. Tal funcionalidade fornece ao usuário dos sistemas de descoberta de especialistas uma melhor compreensão dos resultados. No entanto, apenas algumas evidências especializadas contêm informações de contexto adequadas. Considerando esta questão, desenvolveu-se um método de injeção de contexto onde injeta-se contexto de evidências correlatas em evidências que carecem de contexto.
Dada a proposta de uma nova abordagem para melhorar os resultados dos sistemas de busca de especialistas, com base na hipótese de que um contexto melhora a compreensão de tais resultados, houve a necessidade de validar a hipótese. Como essa avaliação é uma questão subjetiva, adotou-se uma avaliação baseada em feedback recebido de usuários de um sistema de descoberta de especialistas com contexto integrado aos resultados. Conforme demonstrado nas respostas obtidas dos usuários, a maioria considerou benéfico contextualizar os resultados, validando assim a hipótese.

CONSIDERAÇÕES FINAIS

Tanto a proposta do framework Experion quanto o método de injeção de contexto foram submetidos a um experimento qualitativo, onde três especialistas foram entrevistados e solicitados a usar nossa ferramenta e analisar os resultados. Estas entrevistas forneceram várias ideias para trabalhos futuros.
Como trabalhos futuros citam-se: (i) ampliar a fonte de dados utilizada incluindo currículos de outras instituições e também de outras plataformas, além do Lattes, e analisar o desempenho do framework; (ii) como melhorar e analisar a eficácia do método de injeção de contexto proposto; (iii) testar métodos alternativos para calcular a similaridade entre evidências de expertise e definir quais contextos devem ser injetados; (iv) elaborar uma proposta de ranqueamento dos resultados considerando as informações de contexto; (v) estabelecer implementações padrões para os componentes do framework, permitindo fácil utilização de outras fontes de dados além da fonte de dados Lattes e criando bibliotecas compartilhadas de uso comum e; (vi) desenvolver novas formas de descrever os contextos encontrados pelo framework, usando uma forma de descrição mais natural e humana.
Palavras-chave: recuperação de expertise, descoberta de especialista, contextualização.

ABSTRACT

Expertise retrieval automatically identifies and describes the expertise associated with a person. The expertise is generated based on the evidence (publications and activities, for example) associated with the person. After a lengthy review of existing work, which produced a faceted taxonomy for such work and identified several open issues, we focus on improving the user understanding of the results produced by *expert finding* systems. Expert finding deals with, given an expertise of interest, locating candidate experts. Our research hypothesis is that, by including contextual information in the results, the users will better understand them. Current works produce a ranked list of candidate experts. The list contains none or, at most, limited contextualization, focused only on the expertise itself. When finding an expert, other contexts can play an essential role in choosing the right person - for a college discipline position, someone with previous teaching experience is more desirable than someone with no teaching experience. This work introduces **Experion**, a framework that standardizes and allows the contextualization of expertise evidence. It identifies, collects, and analyzes contextual data associated with the evidence, such as where, when, and with whom given expertise evidence has occurred or has been produced. This analysis generates a standardized set of contextual data. Experion, using such data, automatically describes the context for the expertise evidence. We applied Experion to the data from the Lattes platform and developed an expert finding system that contextualizes its results using the framework. Using this system, we promoted qualitative experiments with the users and collected their feedback, which confirmed our hypothesis that contextualization improves the understanding of the results.

Keywords: expertise retrieval, expert finding, contextualization.

LIST OF FIGURES

LIST OF TABLES

CONTENTS

1 INTRODUCTION

Expertise can be loosely defined as the knowledge people acquire through life experiences (BALOG et al., 2012). Expertise retrieval deals with automatically discovering and describing this type of knowledge adequately.

In this chapter, we detail the concepts of *expertise* and *expertise retrieval*. To provide context for this work, we present a brief introduction to the current state of expertise retrieval systems and the motivation for our work. We introduce the objectives, proposal, expected contributions, and research methodology adopted in this work.

1.1 EXPERTISE RETRIEVAL

According to Balog et al. (BALOG et al., 2012), *expertise* is a loosely-defined concept that is not easy to formalize or represent and is 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 knowledge to carry out tasks and solve problems. However, it is difficult for them to express it in a particular, formalized, and complete way that allows other people to know about their expertise. New ways to discover and automatically describe this knowledge adequately and accurately is a valuable and challenging research topic.

One way to perceive tacit knowledge is to analyze the *expertise evidence* associated with a person. *Expertise evidence* is an artifact that contains information related to expertise (BALOG et al., 2012). There are many sources from which these artifacts can be obtained: authored documents (articles, reports), electronic communications, and social networks. Expertise retrieval is finding, extracting, and linking the evidence to specific expertise. Figure 1 introduces this process. In general terms, there are three stages:

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

Expert finding and profiling are two primary applications for expertise retrieval (BALOG et al., 2012). In *expert finding*, given a list of one or more topics of interest, experts related to those topics are located. *Expert profiling* involves building expertise profiles, i.e., descriptions of people's expertise (BALOG; DE RIJKE, 2007).

Understanding and describing the expertise of a person is a time-consuming and complex task. Three factors directly impact manually keeping a description of the expertise of a person:

1. The expertise changes continually, following the activities of the person involved.

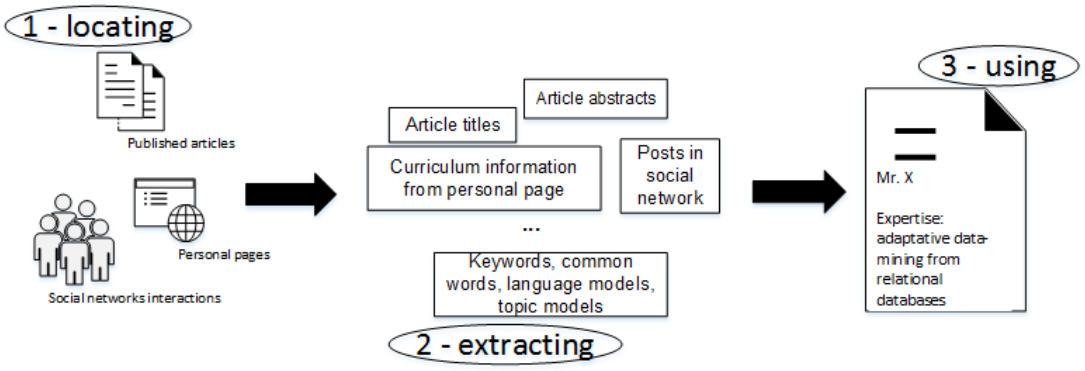


Figure 1 – Expertise retrieval process

2. It requires an understanding of the subjects associated with the expertise of the person describing it.
3. It is not trivial to quantify, i.e., how much expertise specific evidence represents.

Since it is time-consuming and complex to retrieve and describe expertise, automated approaches have become an exciting research topic for many computer science communities in recent years. The information retrieval community (CHEN, H.-H. et al., 2013; HOFMANN et al., 2010; J et al., 2016; BOEVA; BONEVA; TSIPORKOVA, 2014; CABANAC, 2011) researched methods to extract expertise evidence and input from data and clustering. The databases community developed indexation and data structures (GOLLAPALLI; MITRA; GILES, 2013; COHEN; EBEL, 2013; XU, Y. et al., 2012; SERDYUKOV; RODE; HIEMSTRA, 2008). In machine learning, key elements (language models and topic models) (LIU, X. et al., 2014; NAVEED; SIZOV; STAAB, 2011; PAL, 2015; HASHEMI; NESHATI; BEIGY, 2013; FANG; GODAVARTHY, 2014; KUNDU; MANDAL, 2018; LIANG, 2019; DEHGHAN; BIABANI; ABIN, 2019; MUMTAZ; RODRIGUEZ; BENATALLAH, 2019; LIMA; SANTOS, R. L. T., 2022; COCARASCU et al., 2021) were developed.

1.2 MOTIVATION

Existing work in expert finding varies in their techniques to elaborate and represent the expertise associated with a person. Each technique can have its representation. Figure 2 shows three different representations of a researcher's expertise. They could come from the same data using different techniques (term frequency, topic model, and graph - these will be explained in Chapter 2). When presenting their results, existing expert-finding systems lack a representation of the expertise identified or are limited to those representations. Such representations:

1. do not describe the expertise evidence on which the expertise was elaborated and;

2. do not explain how the expertise was obtained or demonstrated - the *context* associated with an expertise.

The context can indicate, for example, where a person has obtained or demonstrated such expertise and with whom. Having a context helps to prevent selecting a professional for a given task with the required expertise in a topic but not in the desired context - for example, selecting a researcher who has never taught before to work as a professor for a given course. To improve existing expert finding systems on these limitations is the key motivation of our work.

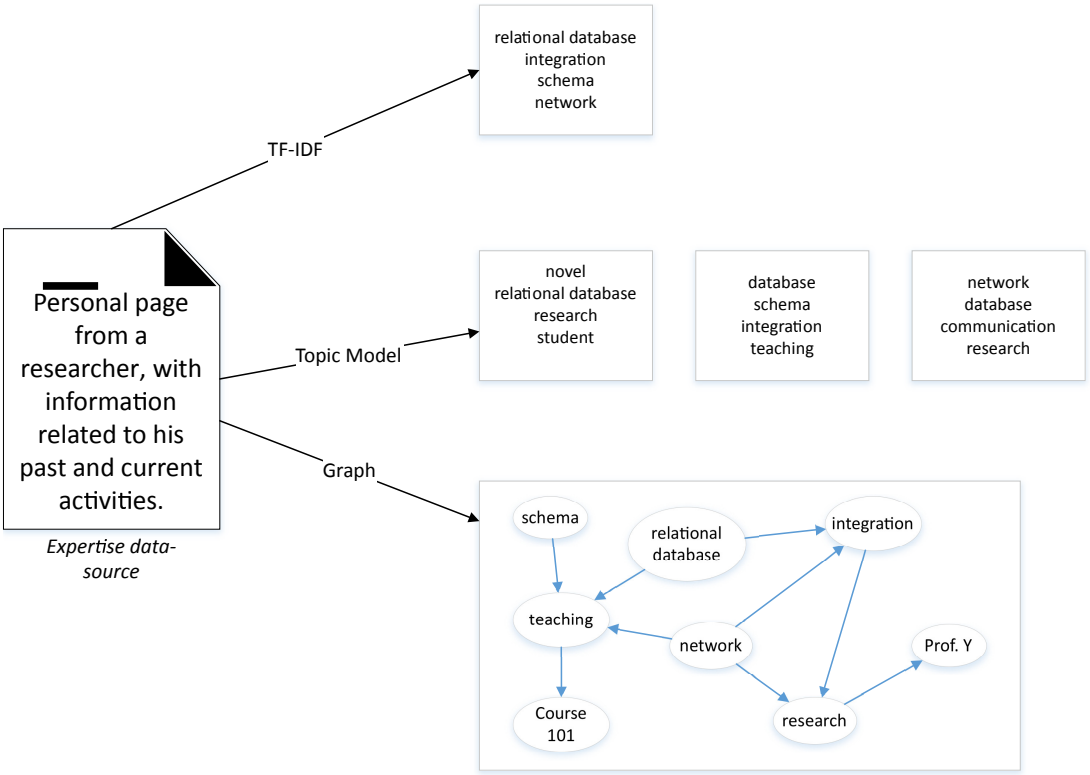


Figure 2 – Expertise representation examples

1.3 OBJECTIVES

This work improves the presentation and understanding of results in expert finding systems. It introduces a new framework called Experion, which provides a contextualized representation of expertise evidence. Expert-finding systems can use this representation to present their results to a user. The following specific objectives are contemplated:

- We propose a standard representation for expertise information (introduced in Section 3.2);
- generate a human-understandable representation of expertise information and include contextual information in expert finding systems by gathering data associated with the expertise evidence (introduced in Section 3.3);

- validate the benefit of more detailed results in expert finding systems by applying the framework over the Lattes platform (Section 4) and promoting experiments (as demonstrated in Section 4.3) and;
- improve the available contextual information associated with expertise evidence with a self-tuning method (described in Chapter 5).

1.4 PROPOSAL AND CONTRIBUTIONS

The main contribution of this work is a novel framework for contextualized expertise representation called *Experion*. The framework includes the following key features and benefits:

- The expertise evidence is structured in a standard and contextualized representation, independent of the original data sources, using the concepts shown in Section 3.2;
- It uses the available data associated with the expertise evidence to elaborate on the contextual information associated with them (demonstrated in Section 4);
- It allows an existing expert finding system to describe its results to the user by:
    1. detailing the expertise evidence considered to elaborate the expertise and;
    2. contextualizing such evidence.

  Section 4.3 introduces an implementation of such a system and the user feedback validating the benefit;
- Identify additional contextual information through a proposed *context injection* method (introduced in Chapter 5).

Compare to previous work in the literature, Experion innovates by proposing a contextualization of how a given Expertise was acquired or demonstrated. Existing work (SATELI et al., 2017) do introduce the idea of contextualization but only for the expertise knowledge (i.e., the concepts associated with it). They do not contextualize environmental aspects such as: (i) where the expertise was obtained; (ii) with whom the Expert interacted; (iii) the purpose of the event that demonstrated expertise - teaching, research, etc. and; (iv) when it happened. Such environmental aspects, in our understanding, play a critical role in understanding the expertise and knowledge of a given person, as well as defining if a candidate expert is a good match for a certain task. For example, an expert with no previous teaching experience may not be the best suited for teaching a course in a University.

Besides Experion, this work also contributes by introducing a novel faceted taxonomy to classify automated expertise retrieval work, resulting from a lengthy analysis of existing work. A survey (GONÇALVES; DORNELES, C. F., 2019) was elaborated, including the novel taxonomy, and published in the ACM Surveys Journal.

1.5 METHODOLOGY

The research methodology adopted in this work included seven phases:

1. executing a literature review;
2. elaborating a survey on expertise retrieval;
3. establishing and developing a proposal;
4. implementing a prototype tool to test and validate the proposal;
5. collecting user feedback on the prototype;
6. promoting further optimization on the proposal, adjusting the prototype tool;
7. collecting additional user feedback and analyzing the results.

Once the literature review was completed, a *survey on expertise retrieval* was elaborated and published (GONÇALVES; DORNELES, C. F., 2019). This survey had a more general approach and intended to confirm the understanding of the state of the art and included a novel taxonomy to classify related work that was developed and introduced in the survey.

The work *proposal* and associated specific objectives were formulated with state of the art understood, including open issues. The proposal was developed, and a *prototype* - an expert finding system - was implemented to validate our proposal and attain the specific objectives associated. An experiment (Section 4.3) was promoted where a group of users was asked to use the developed system. Each user then answered a questionnaire on the impact of the additional expertise information and contextualization, provided by Experion, on the understanding of the results of the expert finding system.

Based on the user feedback from the experiment and our own experience developing the prototype, an additional *optimization* to our proposal was developed, called *context injection*. It uses the contextual data available in the expertise evidence to improve the context of evidence that lack contextual information. This optimization required further developments on the prototype tool, which was then submitted to a qualitative experiment (Chapter 5). We interviewed three experts with more than ten years of experience in their fields, on the user experience using the developed tool and collected their impressions and suggestions. The interviews were semi-structured, with a basic guideline allowing the experts to manipulate the tool freely, and around one hour each. Assistance and further clarifications were provided during the interview

regarding the tool and the framework. Based on this experiment, we attained this thesis proposal and established future work to improve the results.

1.6 STRUCTURE

This work is organized as follows: Chapter 2 presents state of art in Expertise Retrieval, introducing and using the proposed faceted taxonomy as a guideline. The chapter also provides a comparison between related work and details several open issues, which served as the basis to establish this work's proposal.

The *Experion* framework, our main contribution to expertise representation and contextualization, is introduced in Chapter 3. The framework concepts and structure are presented. An application and prototype implementation of the framework, including a description of an experiment performed to collect user feedback, are introduced in Chapter 4.

Chapter 5 describes an additional development proposal to improve the expertise contextualization. This proposal is called *context injection*, and additional experiment results (an interview with experts) are introduced. Chapter 6 concludes our work with an analysis of our results and possible future work.

2 TAXONOMY AND RELATED WORK

Expertise retrieval, as introduced earlier, follows three basic steps (BALOG et al., 2012): (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 and are introduced in detail in Section 2.2. The kind of information searched in these data sources can include:

- titles, keywords, abstracts, and text bodies from articles;
- text from documents stored in knowledge-management systems;
- messages in social networks;
- relations between people and their productions (citations, co-authoring network);
- activities on the web - forums and question & answer sites.

Existing work applies many techniques over the data found to elaborate evidence for expertise retrieval-related tasks. For example, language models (MANNING; RAGHAVAN; SCHÜTZE, 2008a) and topic models (BLEI, 2012a) are built over the text from abstracts and body of articles, documents, web pages, and posts in SNS (Social Networking Service). Language models allow finding people whose documents directly relate to a given query (set of words). Topic models represent the probable expertise associated with a person through a summarized word-based representation of their production. Topic models can also be used to compare people, identifying those with similar expertise. Other works use the relations extracted from SNS and researchers' production (such as co-authoring and citations) to locate those who stand out in a given topic (based on how many people refer to them and how many publications they have). Section 2.4 introduces several examples of how existing work extracts expertise information, and Section 2.5 describes the tasks where this information is used.

After analyzing and identifying the components in expertise retrieval-related work, we elected four key components that compose their 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 following 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 main 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 based on 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.

Following these components, we have created a taxonomy to classify the existing studies. In this chapter, we introduce this taxonomy, which was published (GONÇALVES; DORNELES, C. F., 2019) in the *ACM Computing Surveys* (ACMSUR) and is introduced in Section 2.1. Sections 2.2 through 2.5 describe the taxonomy components in detail while referring the reader to several existing works. Section 2.6 classifies and compares existing work based on the taxonomy proposed. Lastly, Section 2.7 introduces and discusses some open issues we identified during our literature review.

2.1 EXPERTISE RETRIEVAL TAXONOMY

Table 1 presents the elaborated faceted taxonomy. It starts with the **data source** component that contains facets describing data sources used by expertise retrieval systems: (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.

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 **Pre-processing:** 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

The next component, **data extraction**, defines how existing work extract data from the data sources: (i) *Expert composition* examines whether a single or multiple semantic types of data are used to build a person's expertise; (ii) *Pre-processing* 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 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 expertise knowledge is formulated: (i) *Method* defines the particular way the expertise is represented; (ii) *Temporal support* is concerned with whether the time is included in the expertise analysis - a person's expertise can vary during his/her lifetime - outdated evidence of expertise may have to be rejected; and; (iii) *Semantic support* is concerned with whether the work uses semantic tools in its processes.

The last component, **Application**, contains facets related to the Application that requires the expertise retrieval process. Currently, it has a single facet called *Task*, but other facets can be added later.

2.2 DATA SOURCE

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

2.2.1 Format

The *Format* facet follows the well-known unstructured, semi-structured and structured classification (ABITEBOUL; BUNEMAN; SUCIU, 2000). **Unstructured data** includes data sources where the semantics associated with the data is null or minimal. An account is taken here of any data source that does not impose semantics on its data. These data sources might include e-mails messages (VAN GYSEL; RIJKE; WORRING, 2016; ALARFAJ; KRUSCHWITZ; FOX, C., 2013; FANG; SI; MATHUR, 2010), scientific articles (LI, C. et al., 2015; KAYA; ALHAJJ, 2014; RYBAK; BALOG; NØRVÅG, 2014), wiki contents (PAL, 2015; OSBORNE; MOTTA; MULHOLLAND, 2013), forums (PAL, 2015; LIU, Jingyuan et al., 2014), question & answer sites (KUMAR, V.; PEDANEKAR, 2016; CHENG et al., 2015), social network posts (LI, C. et al., 2015; LIU, D. et al., 2013), web-pages (PAL, 2015; LI, C. et al., 2015; DAVOODI; KIANMEHR; AFSHARCHI, 2013) and knowledge management systems (KMS) (VERTOMMEN et al., 2008; YANG, K.-W.; HUH, 2008). When extracting data, unstructured data sources raise some challenges. If we find the name of someone at a prestigious university, this data can be subject to several interpretations: (i) the person works (or has worked) at this university; (ii) co-authored a paper with someone else from this university. It could also be that he studied there, or it may even just be a citation of places where he 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 severe since our objective is to find evidence of expertise so that a person can be profiled. In an article by Zhan et al. (ZHAN et al., 2011), for example, an 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 authorship order does not always indicate to what extent a person is an expert on the article's subject.

In **semi-structured data sources**, there is support in their formats for introducing semantics to the data. The two primary examples of semi-structured data are XML and JSON documents.

In expertise retrieval, there are several cases where semi-structured sources are used. For example, bibliographic databases publish their contents as XML files, such as DBLP (LEY, 2009) and CiteSeer (LI, C. et al., 2015). DBLP and CiteSeer gather information from several sources and provide an integrated overview and search mechanism for this data. Another example is AMiner (TANG, Jie, 2016), an academic search engine and mining system that exports its data through an API (JSON) and semi-structured text files[1]. Stack Overflow also publishes its data as XML files[2]. Semi-structured 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 semi-structured formats such as BibteX.

Some studies combine semi-structured and unstructured data sources, Li et al. (LI, C. et al., 2015) and Fang et al. (FANG; GODAVARTHY, 2014), 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 retrieved from semi-structured data sources. Unstructured data sources (the full text of the articles) were then queried to extract further information about the related expertise.

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

To our knowledge, no published work explicitly uses structured data sources. That does not mean that these data sources cannot be used for expertise retrieval work. Relational databases of institutions (such as universities or research centers) can include valuable information on a person's academic output and activities (LIU, Ping; CURSON; DEW, 2002). Retrieving and using this kind of data can significantly assist in designing an expertise profile.

[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/

2.2.2 Accessibility

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

Public unrestricted 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 primarily unstructured and semi-structured, and web pages are examples. In the context of expertise retrieval, there are personal, project, and institutional pages. Mailing list archives are another example of this level of accessibility. Data sources available in the Surface Web (BERGMAN, 2001) are examples of public unrestricted data sources as well - and include, for instance, public wikis such as Wikipedia[6]. Examples of public unrestricted semi-structured data sources include DBLP, CiteSeerX, and the Stack Exchange network. DBLP publishes a *dump* of its 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)(LAGOZE; VAN DE SOMPEL, 2001) interface. The Stack Exchange network publishes its data as XML files through the Internet Archive project[7].

Public restricted sources (mainly in the Deep Web) that provide access to their information entail one or more of the following:

- licensing costs;
- 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 they are not computer-friendly. Their policies, in some cases, explicitly state that computer-based data harvesting is forbidden[8]. Other publishers, like Elsevier, provide APIs to access their records but limit the results per query instance. These publishers also restrict the available data - only subscribers can access the full text of the articles indexed for them, while non-subscribers may access only meta-data.

The Lattes Platform [9] standardizes the curriculum format for Brazilian researchers.

[6] http://www.wikipedia.org
[7] https://archive.org/details/stackexchange
[8] http://librarians.acm.org/policies
[9] http://lattes.cnpq.br/web/plataforma-lattes

It is a central directory allowing Brazil researchers 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 classified as *public restricted data sources* is social network services (SNS), such as Facebook, ResearchGate[10], and LinkedIn[11]. Facebook, for example, provides an API (called *Graph 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 (CHAIWANAROM; LURSINSAP, 2015; RIBEIRO et al., 2015; RYBAK; BALOG; NØRVÅG, 2014; LIN et al., 2013), 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. For example, Fang et al. (FANG; GODAVARTHY, 2014) use data from AMiner together with Google Scholar[14] crawled abstracts.

Lastly, **private data sources** do not provide public access. These sources are only accessible through internal networks in institutions/corporations. External users may not even know about their existence. Examples can be cited, such as private KMS (YANG, K.-W.; HUH, 2008), private SNS (PAL, 2015), and micro-blogs (LIU, D. et al., 2013). Although not found in any of the studies reviewed, private wiki and HU (Human Resources) systems could equally be a private source of expert information. Intranet communications and e-mails (KARIMZADEHGAN; WHITE; RICHARDSON, 2009; ZHU, J. et al., 2005) are possible data sources for expertise and are classified as private data sources as well.

2.2.3 View

The third facet of the data source component is how the data can be viewed. Three classification types are proposed for the existing data sources: plain text, communications, and dataset.

Plain-text views are found in sources composed of unstructured documents, requiring 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 (DAVOODI; KIANMEHR; AFSHARCHI, 2013;

[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/

BERENDSEN et al., 2013; PAL, 2015; BALOG; AZZOPARDI; RIJKE, 2009; TANG, Jie et al., 2008). Collections of documents in a corporation or institution are examples of plain-text data sources (SERDYUKOV et al., 2011; YANG, K.-W.; HUH, 2008; FAZELZARANDI; FOX, M. S., 2011).

Communications views can be found in sources that represent message exchange. They include: instant messaging(ZHU, J. et al., 2005), e-mails, and mailing lists(CAMPBELL et al., 2003; DOM et al., 2003; BALOG; RIJKE, 2006; BALOG; AZZOPARDI; RIJKE, 2006; XU, Y. et al., 2012; SERDYUKOV; RODE; HIEMSTRA, 2008; FANG; SI; MATHUR, 2010; ZHU, Jianhan et al., 2010; BALOG; AZZOPARDI; RIJKE, 2009), Web forums(BERENDSEN et al., 2013; PAL, 2015; BALOG; AZZOPARDI; RIJKE, 2009) and question & answer sites(CHENG et al., 2015; BUDALAKOTI; DEANGELIS; BARBER, 2009; LIU, Jingyuan et al., 2014; BHANU; CHANDRA, 2016). SNS (Social Network Systems)(LI, Y.; TANG, J., 2008; LIU, D. et al., 2013; YANG, Chen et al., 2014; YANG, C. et al., 2015) 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 an SNS has clear information regarding personal relations (friends, acquaintances), it can be seen as a communication system with additional information available from an expertise retrieval perspective. This information includes social interactions and relations.

A *dataset* view involves providing data in a standardized and queryable format. It includes information systems, databases, and files in structured/semi-structured formats. The indexers of scientific 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(CHAIWANAROM; LURSINSAP, 2015; FANG; GODAVARTHY, 2014; SHI et al., 2012; ZHENG et al., 2011). Wiki and KMS systems also provide dataset views through meta-data associated with their web pages, such as author and abstract(YANG, K.-W.; HUH, 2008; PAL, 2015; DAVOODI; KIANMEHR; AFSHARCHI, 2013). Bibliographic databases (DBLP, CiteSeer[15], AMiner[16], ScholarMiner) are also examples of data sources with a dataset view(BOLELLI; ERTEKIN; GILES, 2009; CABANAC, 2011; CHEN, H.-H. et al., 2011; COHEN; EBEL, 2013; DAVOODI; KIANMEHR; AFSHARCHI, 2013; FANG; GODAVARTHY, 2014; KOU et al., 2015; LI, J. et al., 2014; LI, C. et al., 2015, 2015; MANGARAVITE et al., 2016; RYBAK; BALOG; NØRVÅG, 2014; SHI et al., 2012; STEYVERS et al., 2004; TANG, Jie et al., 2012). DBLP provides an XML *dump*[17]

[15] http://citeseerx.ist.psu.edu
[16] http://aminer.org/
[17] http://dblp.uni-trier.de/xml/

of its data, and 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 a dataset view. The set of published pages is a plain text view, while their associated meta-data (internal to the wiki system) is a dataset view. The same happens with SNS - if a system only accesses public pages from an SNS that originates from links in other pages, it will regard it as plain text. However, it will see the same source from a communications view if it examines the relations between people through the SNS interface/API.

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

2.3 DATA EXTRACTION

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

2.3.1 Expert composition

Expert composition defines the extent to which more than one semantic data type 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 the body of scientific articles can be considered the same kind of information, i.e., keywords. On the other hand, although co-authors' names and keywords are the same kinds 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 researchers can be represented as a set of words extracted from their papers. *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 studies that introduce expert composition based on terms extracted from linked documents are simple compositions (BALOG; AZZOPARDI; RIJKE, 2009; GOLLAPALLI; MITRA; GILES, 2012; MANGARAVITE et al., 2016; KAWAMAE, 2010; LI,

[18] http://csxstatic.ist.psu.edu/about/data
[19] http://doc.aminer.org/en/latest/

C. et al., 2015; MIMNO; MCCALLUM, 2007; TANG, Jie et al., 2008; VERTOMMEN et al., 2008; JOHRI; ROTH; TU, 2010; CABANAC, 2011; DAVOODI; KIANMEHR; AFSHARCHI, 2013; J et al., 2016; BALOG; RIJKE, 2007; COHEN; EBEL, 2013; RIBEIRO et al., 2015). Some works (BOLELLI; ERTEKIN; GILES, 2009; DAUD, 2012; NAVEED; SIZOV; STAAB, 2011) 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 (co-authoring, citations, and social relations) (KOH; DOBBIE, 2012), while others introduce complex compositions based on co-authoring and temporal factors (HUANG et al., 2014; LI, J. et al., 2014). Related entities, such as the venues where an author's work has been published (KOU et al., 2015), are also used. Combining concepts extracted from the documents and co-authoring information (XU, Y. et al., 2012) drawn on to represent an expert is another approach used in literature. Parada et al.(PARADA et al., 2013) combine different features to represent the range of the researcher's interests.

2.3.2 Pre-processing

Some studies employ pre-processing techniques for the extracted data to improve expertise retrieval. Three categories are proposed to classify existing work:

- *None* for works that do not apply pre-processing techniques;
- *Word removal* for those that remove unnecessary words from the text;
- *Text transformation* for those that transform the text.

Stop word removal is the most common example of the word removal category. 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 (NAVEED; SIZOV; STAAB, 2011; KAWAMAE, 2010; JIANG; LI, X.; MENG, 2014; DAUD et al., 2009; KAWAMAE, 2010; TANG, Jie et al., 2011). Some also remove punctuation and numbers(YANG, Z.; HONG; DAVISON, 2013; BOLELLI; ERTEKIN; GILES, 2009). *Stemming* (SINGH; GUPTA, 2016) reduces words to their root form and is the most used technique that performs *text transformation*. Johri et al.(JOHRI; ROTH; TU, 2010) normalize authors' names while analyzing their publications.

2.3.3 Retrieval

When extracting data, the existing studies retrieve all 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 studies that execute a complete retrieval do not provide results based on a given input/query but produce a result that can be

browsed and analyzed by a user. They include works that: are designed to build researcher profiles (TANG, Jie et al., 2008; CHAIWANAROM; LURSINSAP, 2015; TANG, Jie et al., 2012; FANG; GODAVARTHY, 2014; GOLLAPALLI; MITRA; GILES, 2012; RYBAK; BALOG; NØRVÅG, 2014; BALOG; RIJKE, 2007), or conduct a co-author network analysis(LI, J. et al., 2014; CHEN, H.-H. et al., 2011; XU, Y. et al., 2012; COHEN; EBEL, 2013; YANG, C. et al., 2015) and research topic analysis(KAWAMAE, 2010; TU et al., 2010; JOHRI; ROTH; TU, 2010; LI, C. et al., 2015; BOLELLI; ERTEKIN; GILES, 2009; KOU et al., 2015).

Focused retrieval work create a query with user input to define the subset of the data that needs to be extracted(SMIRNOVA; BALOG, 2011; PAL, 2015; FANG; SI; MATHUR, 2010; DENG et al., 2012; TANG, Jie et al., 2010; HASHEMI; NESHATI; BEIGY, 2013; LI, Y.; TANG, J., 2008; LIU, D. et al., 2013; DENG et al., 2012; SERDYUKOV; RODE; HIEMSTRA, 2008; GOLLAPALLI; MITRA; GILES, 2013; MACDONALD; OUNIS, 2009) . They can be adapted to existing standard search engines with little effort(BALOG; AZZOPARDI; RIJKE, 2009). Most are deployed in expert finding.

2.4 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*.

2.4.1 Method

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

Term Frequency uses the frequency of terms in a document to define or retrieve the expertise related to a person (BALOG; RIJKE, 2007; BOEVA; BONEVA; TSIPORKOVA, 2014; CABANAC, 2011; COHEN; EBEL, 2013; DAVOODI; KIANMEHR; AFSHARCHI, 2013; DUONG; NGUYEN; JO, G. S., 2010; J et al., 2016; KUMAR, A.; JAIN, 2010; VERTOMMEN et al., 2008). Its primary aim is to consider how many times a term appears in a given document to assess its relevance.

The term frequencies can be used to build vectors (VERTOMMEN et al., 2008) that represent an author(HOFMANN et al., 2010; COHEN; EBEL, 2013; GOLLAPALLI; MITRA; GILES, 2012; J et al., 2016; BOEVA; BONEVA; TSIPORKOVA, 2014; CABANAC, 2011) or a document(THO; HUI; FONG, 2003; CHEN, H.-H. et al., 2013; LIU,

D. et al., 2013). 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(CABANAC, 2011) 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(BOEVA; BONEVA; TSIPORKOVA, 2014) or collaboration recommendation(COHEN; EBEL, 2013)).

According to Manning et al. (MANNING; RAGHAVAN; SCHÜTZE, 2008b), a language model is a *"function that puts a probability measure over strings drawn from some vocabulary."* It is formed based on an existing text by analyzing how frequently specific terms (a word or group of words) appear. Each term is assigned a relative frequency 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 him/her (MANGARAVITE; SANTOS, R. L., 2016; YANG, C. et al., 2015; FANG; GODAVARTHY, 2014; BALOG; AZZOPARDI; RIJKE, 2009; BALOG; DE RIJKE, 2007; BALOG; RIJKE, 2007; SMIRNOVA; BALOG, 2011; HASHEMI; NESHATI; BEIGY, 2013).

Topic Models(BLEI, 2012b) are probability distributions regarding a given document's topics. 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 through its meta-data, for example, a list of related categories. When seeking the topics of a given author, the categories related to each document 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 (BLEI, 2012b), topics are defined as clusters of related words obtained through statistical analysis. Each topic is represented by a set of words, defined through sampling methods such as those of Gibbs (GRIFFITHS; STEYVERS, 2004).

Topic models do not limit themselves to modeling topics arising from documents. Several studies introduce *hidden variables*, i.e., probability distributions regarding the domain's features. A hidden variable can represent (i) the probability of an author writing about a given topic based on the topics of his related documents; (ii) how likely a given topic will feature in a conference based on articles from previous editions. 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, Topic models were found to be the most common expertise evidence extraction

technique in the reviewed studies.

There are several approaches to solving the problem of topic modeling. In the context of expertise retrieval, a well-known approach is LDA (Latent Dirichlet Allocation), introduced by Blei et al. (BLEI; NG; JORDAN, 2003). It designs a generative probabilistic model for data collection through a three-level hierarchical Bayesian model applicable to text corpora. The *Author-Topic-Model*, outlined by Rosen-Zvi et al.(STEYVERS et al., 2004), also models the topic distribution for the authors concerned, i.e., by determining which topics are shared by each author. Other studies (CHA et al., 2015; JAMEEL; LAM, 2013a, 2013b; CHAIWANAROM; LURSINSAP, 2015; PAL, 2015; JIANG; LI, X.; MENG, 2014; LIN et al., 2013; DU et al., 2015; KOU et al., 2015; WANG, X.; ZHAI; ROTH, 2013; MOU et al., 2015) extend the Author-Topic-Model by including additional features such as venues (YANG, Z.; HONG; DAVISON, 2013; CHEN, X.; ZHOU, M.; CARIN, 2012), document citations (KATARIA et al., 2011; TANG, Jie et al., 2011), pre-existing supervised document subject classification (MOU et al., 2015) or cooperation information between authors to improve the efficiency of topic discovery (GAO et al., 2017). Xie et al.(XIE et al., 2016) introduced a topic model that covers social interactions and relationships in social networks when building an expert's topic model and ranking his/her expertise on the desired topic.

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(DENG et al., 2012) or by using transformed data that are generated through another technique (topic and language models(CHAIWANAROM; LURSINSAP, 2015; YANG, C. et al., 2015; LIN et al., 2013)). Once the graph has been generated, specialized methods are employed to extract the required information. *Page-rank-like* methods such as Random Walk(SERDYUKOV; RODE; HIEMSTRA, 2008; GOLLAPALLI; MITRA; GILES, 2013) are used for collaboration recommendation and expertise retrieval (i.e., expert ranking). Other studies(CHEN, H.-H. et al., 2011) use the graph's structure to calculate similarities between entities (such as authors) and recommend collaborations.

Some works (LIN et al., 2013; PENG et al., 2013) use weighted graphs to represent, for example, co-authorship networks(PENG et al., 2013). Other examples of graph-based techniques include a) author profiling based on co-authors (HOANG; KHOA; PHUC, 2013), b) finding the most probable author for a given topic (SERDYUKOV; RODE; HIEMSTRA, 2008; LIN et al., 2013) and the closest people to a given person (YANG, C. et al., 2015; CHAIWANAROM; LURSINSAP, 2015) or c) clustering people in terms of their expertise (J et al., 2016; KOU et al., 2015). Kong et al.(KONG et al., 2017) and Xie et al.(XIE et al., 2016) combine topic model comparisons between authors with a random-walk technique within a collaboration graph to suggest collaborations. Robertie et al.(LA ROBERTIE et al., 2017) set out 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 also been proposed (LIU, Z.; LI, K.; QU, 2017) as a tool for finding users to answer questions in CQAs (Community Question Answer) sites.

This category includes studies that use alternative techniques. For example, Punnarut et al.(PUNNARUT; SRIHAREE, 2010) created a researcher profile based on an ontology that extracted terms from documents and matched them with a previously defined list of skills. Latif et al.(LATIF; AFZAL; TOCHTERMANN, 2010) uses LOD (Linked Open Data) (BIZER et al., 2008), which is available on the web, to build researchers' profiles. Linked Open Data is a method based on standard Web technologies such as HTTP, RDF, and URIs. It is employed to publish data so that computers can automatically interpret and handle it.

Fang et al.(FANG; SI; MATHUR, 2010) 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.(MACDONALD; OUNIS, 2009) proposed using the voting model to rank experts from an expert search result. Ban et al.(BAN; LIU, L., 2016) sought to combine graph techniques (using a citation network) and introduced a customized VSM (which includes the location where the terms appear in an article so that they can be weighted) as a way to find experts.

2.4.2 Temporal support

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

Suppose, for instance, there is a need for an expert in *GIS databases*. Two researchers are selected, based on their output, *ResearcherA*, and *ResearcherB*. ResearcherA published many works ten years ago, but in recent years has focused his research on distributed transactions. ResearcherB only started to publish papers on GIS databases 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 decide based on what activity is required from the expert. For lecturing undergraduate students about GIS databases, either ResearcherA or ResearcherB could be invited. However, regarding integrating a new research project, ResearcherB should be preferred since he will probably be more interested in this than ResearcherA, who has recently changed his research field.

Defining time's effects on the expertise evidence is not a trivial issue. Many studies (CHAIWANAROM; LURSINSAP, 2015; JAMEEL; LAM, 2013a; NAVEED; SIZOV; STAAB, 2011; HE et al., 2009; ZEHNALOVA et al., 2012; HASHEMI; NESHATI; BEIGY, 2013; LI, Y.; TANG, J., 2008; COHEN; EBEL, 2013; FANG; GODAVARTHY, 2014; RY-

BAK; BALOG; NØRVÅG, 2014; JO, Y.; HOPCROFT; LAGOZE, 2011; KAWAMAE, 2012; BOLELLI; ERTEKIN; GILES, 2009; WANG, X.; ZHAI; ROTH, 2013; DAUD, 2012; XU, S. et al., 2014) include time in their analysis and the extraction of expertise evidence. Three possible ways on *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*.

In **time slices** approaches, the evolving pattern of expertise is analyzed in *slices* of time, such as a year, for example, where each slice can be influenced by evidence from previous slices. For example, Chaiwanarom et al.(CHAIWANAROM; LURSINSAP, 2015) analyzed the evolving expertise of a researcher using his/her topics of interest over 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.(FANG; GODAVARTHY, 2014) and Rybak et al.(RYBAK; BALOG; NØRVÅG, 2014) 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.(KONG et al., 2017) build per-year Topic Models (LDA) for the author's output by analyzing their active research interests over the years.

Neshati et al.(HASHEMI; NESHATI; BEIGY, 2013) examined the question of research longevity (as the number of years) when estimating how strong the relationship between an author and published paper topics when he/she is a co-author. Li et al.(LI, Y.; TANG, J., 2008) used a time-partitioned random walk in a graph to analyze evolving expertise in a social network. Bolelli et al.(BOLELLI; ERTEKIN; GILES, 2009) included the time factor in their topic model (S-ATM - Segmented Author-Topic Model) when they analyze the topic evolution per time unit (year), in which previous years have a decaying influence on the current year. Daud et al.(DAUD, 2012) proposed TAT (Temporal-Author-Topic), which introduces a similar idea.

Jin et al.(JIN et al., 2017) analyzed the number of publications per topic and year to discover changes and tendencies in the expert's interests. Neshati et al.(NESHATI; FALLAHNEJAD; BEIGY, 2017) analyzed the evolving topic model of experts based on Q&A sites. They introduced four features that affect the topic transitions:

1. topic similarity - users usually change between similar topics;
2. emerging topics - users tend to prefer emerging topics;
3. user behavior - how common it is for users to explore and change topics of interest;
4. topic transition - determining which topic changes are most common.

In **continuous** approaches, there is no need for pre-defined time slices when the expertise evolution is being analyzed. For example, Jameel et al. (JAMEEL; LAM, 2013a) 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 et al. (NAVEED; SIZOV; STAAB, 2011) included absolute timestamps in their topic model (ATTention). Kawame et al.(KAWAMAE, 2012) also included timestamps in their Theme Chronicle Model, and defined the concepts of stable and dynamic topics.

He et al. (HE et al., 2009) analyzed topic evolution through citations between papers by considering Topic Model, which models documents as two independently generated parts: an *inherited* and an *autonomous* part. The former is the outcome of previous work (based on the citations found). Wang et al.(WANG, X.; ZHAI; ROTH, 2013) also included the time when analyzing topic evolution in their Citation-LDA Topic model. Jo et al.(JO, Y.; HOPCROFT; LAGOZE, 2011) analyzed a collection of documents in chronological order and established topic evolution. Estimating the future expertise of users in CQAs sites, including the transition probability to a new topic, has also been researched (NESHATI; FALLAHNEJAD; BEIGY, 2017).

Zehnalova et al.(ZEHNALOVA et al., 2012) devised a *forgetting function* to analyze an author's topic evolution over time. Cohen et al.(COHEN; EBEL, 2013) examined the time elapsed when analyzing authors' collaborations in co-authoring networking. Xu et al. (XU, S. et al., 2014) introduced the Author-Topic over Time (AToT) model. This topic model includes a timestamp associated with the topics used to design an author's interest model and its changes over time. Xie et al.(XIE et al., 2016) 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 microblogs, timestamps, and Internet usage.

2.4.3 Semantic support

The last facet in the expertise representation component of the taxonomy classifies existing work 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 which rely on ontologies, some require an ontology prepared in advance (TANG, Jie et al., 2008; LIU, P.; LIU, K.; LIU, J., 2007; RYBAK; BALOG; NØRVÅG, 2014), while others construct one during their processes (XU, Y. et al., 2012; FAZEL-ZARANDI; FOX, M. S., 2011; KAMSIANG; SENIVONGSE, 2014; PUNNARUT; SRIHAREE, 2010). Those that rely on lexical databases use them to build ontologies that are based on word relations (XU, Y. et al., 2012; KAMSIANG; SENIVONGSE, 2014) or to overcome problems regarding the usage of terms in documents (such as synonyms, hypernym, and hyponym) by finding equivalent words(BOEVA; BONEVA; TSIPORKOVA, 2014). Two examples of knowledge bases used by researchers are DBpedia (OSBORNE; MOTTA; MULHOLLAND, 2013) and Wikipedia (DAVOODI; KIAN-

MEHR; AFSHARCHI, 2013; CHEN, H.-H. et al., 2013).

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.(DAVOODI; KIANMEHR; AFSHARCHI, 2013) 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.

2.5 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. Each task is discussed with information about the particular features of related work. In each Application, we elected representative work to introduce a more detailed discussion.

2.5.1 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 approaches for expert finding in the literature: (i) compiling a specialist index based on expertise-related information (LIN et al., 2013; CHEN, H.-H. et al., 2013; THO; HUI; FONG, 2003; LI, Y.; TANG, J., 2008; WANG, J. et al., 2012; TU et al., 2010; TANG, Jie et al., 2010, 2011, 2008; HASHEMI; NESHATI; BEIGY, 2013; PENG et al., 2013; DENG et al., 2012); and (ii) using traditional indices (such as an inverted index) to locate documents related to given expertise and employ expert finding methods based on the results (PAL, 2015; FANG; SI; MATHUR, 2010; LIU, X. et al., 2014; LIU, D. et al., 2013). Although most approaches look for a single expert, there are studies in the literature that focus on finding groups of experts as well (NESHATI; BEIGY; HIEMSTRA, 2014; LIANG; RIJKE, 2016).

Current studies in the field adopt several approaches to finding experts. Many use document-centric methods, such as: (i) using SVM (support vector machine) to find, represent and search for experts, given keywords of interest (LI, Y.; TANG, J., 2008; CHEN, H.-H. 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 (LIU, X. et al., 2014; WANG, J. et al., 2012; TU et al., 2010; LIN et al., 2013; CHEN, H.-H. et al., 2013; PAL, 2015; TANG, Jie et al., 2010, 2011, 2008; MIMNO; MCCALLUM, 2007; LIANG;

RIJKE, 2016); (iii) representing expertise through ontologies and using them to search for experts(PUNNARUT; 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; HUI; FONG, 2003). Some studies use alternative information sources: (i) bibliographic network information(HASHEMI; NESHATI; BEIGY, 2013; PENG et al., 2013); and (ii) online activities such as posts in CQAs (Community Question Answer) (LIU, D. et al., 2013; BHANU; CHANDRA, 2016), blogs (LI, Y. et al., 2012) and SNS (Social Network Systems) (NESHATI et al., 2014). There are also specific approaches, such as converting tag-based classification of questions in CQAs to topic model representations to find relevant experts for a given question (DARGAHI NOBARI; SOTUDEH GHAREBAGH; NESHATI, 2017) or using geo-tagged information to locate experts associated with specific places (LI, W.; EICKHOFF; VRIES, 2016).

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 he/she can write about different combinations of topics for each publication (MIMNO; MCCALLUM, 2007). With the aid of bibliographic network information, central authors (well-cited or with many co-authorships) can be found. This centrality can be used as an indication of expertise (PENG et al., 2013). Citation counts related to the longevity of research topics (same topic present over extended periods) are also regarded as an indication of expertise (HASHEMI; NESHATI; BEIGY, 2013). Some studies combine expertise evidence and social network relationships (co-authorship or online community meta-data, for example) to find experts (LIN et al., 2013; DENG et al., 2012; LI, Y. et al., 2012; LIU, D. et al., 2013).

Domain-specific approaches, such as examining how difficult questions are answered in CQAs sites, indicate that there is expertise as well (BHANU; CHANDRA, 2016). Machine-learning approaches to locate *future* experts in CQAs sites, combining features from several types (textual, behavioral, and time-aware), have also been proposed (DIJK; TSAGKIAS; RIJKE, 2015).

CSSeer (CHEN, H.-H. et al., 2013) locates experts using data available in Citeseer, supported by data extracted from Wikipedia. First, it extracts keyphrases (bi, tri, and quadgrams) from Wikipedia pages about computer science, statistics, and mathematics. The keyphrases which appear at least three times in the collection of documents from Citeseer are considered *keyphrase candidates*, and the documents are indexed based on these keyphrases. To locate an expert, given a query input (set of words), it locates all authors from documents textually relevant (based on the keyphrases), 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. (PAL, 2015). They crawled data on 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. They extract several features to classify the documents using a question modeler (linear SVM classifier) and a self-developed algorithm (DOCSENSE). These include content features (topic distribution, hashtags, referenced entities); social features (is it a reply for a question, a recommendation, something being shared); processed features (DOCSENSE features such as if the document is non-relevant for expertise analysis if it is a duplicate from other documents) and; 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) is separately indexed so that they can 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). It accounts for the proximity between query words in the retrieved documents to rank them. Lastly, they use GMM (Gaussian Mixture Model) to discard retrieved documents whose topic distribution is irrelevant 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. After this, the expertise score for each source is calculated. Last, an SVM rank aggregation algorithm combines the various sources' scores to calculate the final expertise score.

2.5.2 Expert ranking

Given the expertise of interest and several candidate experts, expert ranking (GOLLAPALLI; MITRA; GILES, 2013; SMIRNOVA; BALOG, 2011; DENG et al., 2012; YANG, Z.; HONG; DAVISON, 2013; TANG, Jie et al., 2008; MACDONALD; OUNIS, 2009; TU et al., 2010; ZHAO et al., 2016) aims to rank these candidates. Most expertise retrieval methods include expert ranking as part of their process since having a ranked list makes more sense than an unordered list (LIU, Jingyuan et al., 2014).

Most studies adopt *graph-based* approaches, such as random-walk, to rank retrieved experts (YANG, Z.; HONG; DAVISON, 2013; GOLLAPALLI; MITRA; GILES, 2013; TANG, Jie et al., 2008). Alternative approaches include: (i) applying neural networks in combination with random-walk methods to rank experts (ZHAO et al., 2016); and (ii) using a regularization framework applied to a heterogeneous network comprising authors and documents (DENG et al., 2012).

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 (TU et al., 2010). The voting model, a ranking technique from the area of *data fusion*, is also used (MACDONALD; OUNIS, 2009). User-centric approaches, such as ranking the experts

based on knowledge gain and easiness of access, were also found in the literature (SMIRNOVA; BALOG, 2011).

Gollapalli et al. (GOLLAPALLI; MITRA; GILES, 2013) introduce an expert ranking method based on two techniques. One technique uses a self-developed ADT (author-document-topic) model (a weighted tri-partite 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 also introduced (using pre-calculated associated weights such as LDA). The authors associated with the documents are introduced in the graph, as well as nodes representing the initial query. 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 paths weights instead of adding them.

In the PageRank-based approach, an initial set of documents is retrieved given a query. A graph is built using these documents and their associated authors. Related documents and authors (through citations, for example) 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.

2.5.3 Expert profiling

Expert profiling provides a virtual representation of a person based on their expertise(NAVEED; SIZOV; STAAB, 2011; LI, Y.; TANG, J., 2008, 2008; BALOG; DE RIJKE, 2007; BALOG; AZZOPARDI; RIJKE, 2009; ZEHNALOVA et al., 2012; PUNNARUT; SRIHAREE, 2010; LATIF; AFZAL; TOCHTERMANN, 2010; LI, Y. et al., 2012). An essential factor in expert profiling is deciding which elements/features are important to include in a person's profile(LATIF; AFZAL; TOCHTERMANN, 2010; BALOG; DE RIJKE, 2007). Some studies that focus on expert finding generate profiles during their procedures and can be used for expert profiling (BALOG; AZZOPARDI; RIJKE, 2009).

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 it will be up to the user to deduce the meaning. Some studies rely on external support, such as ontologies, to build human-understandable profiles (PUNNARUT; SRIHAREE, 2010).

There are many approaches to forming expert profiles. They vary both in their techniques and data sources: (i) some use online information such as intranet web pages (ZHU, J. et al., 2005), Linked Open Data (LATIF; AFZAL; TOCHTERMANN, 2010) or topics in online communities data (LI, Y. et al., 2012); (ii) other studies refer to social relations to help build the profile, such as propagating expertise (LI, Y.; TANG,

J., 2008; J et al., 2016) or inferring expertise (HOANG; KHOA; PHUC, 2013) through related authors; and (iii) some works analyze expertise in the context of temporal evolution and demonstrate how the expertise evolves (NAVEED; SIZOV; STAAB, 2011; ZEHNALOVA et al., 2012; LI, Y.; TANG, J., 2008; FANG; GODAVARTHY, 2014; RYBAK; BALOG; NØRVÅG, 2014; DAUD, 2012).

Fang et al. (FANG; GODAVARTHY, 2014) introduces an interesting application. It calculates how probable it is that, given an expert, he/she will stay in his/her current areas of expertise or change to new areas. They analyze how the publications associated with an expert vary their topics over time. They use the associated keywords to define the topics associated with a document. All the abstracts of documents associated with a given keyword are then analyzed to define the topic model (set of words) associated with the topic.

Based on the volume of publications associated with each topic in each year, they introduce a probabilistic model to calculate if an expert: (i) will stay in his/her current research areas or; (ii) will migrate to new areas. Three features are considered:

1. Is it common for the expert to change areas based on past years?
2. How similar is a new area to the expert's current areas?
3. 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 this set of topics (represented by the topic model words associated with them) and their associated probabilities, previously calculated. Given a query topic, the PLM calculates how probable an expert will be to research the topic.

Author2Vec (J et al., 2016) is an unsupervised machine learning approach to estimate an author's representation as a vector of embeddings extracted from his/her papers using Paragraph2Vec. A neural language model uses the distance and angular similarity between vectors (the author vector and the paper vector) to learn the author's vector representation. The neural network is supplied with 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.

ScholarLens (SATELI et al., 2017) is a platform that, given a set of articles from a researcher, identifies and extracts named entities (using NLP methods) and, using DBPedia as support, elaborates a knowledge database representing the researcher knowledge. The profiles are built using RDF, and the competencies (expertise) are modeled using the IntelLEO ontology, allowing SPARQL queries.

2.5.4 Expert clustering

Automated expertise retrieval makes it possible to cluster people with similar expertise. Some studies use graphs and similarity metrics for this task. The similarity is calculated from the contents of associated documents (BOEVA; BONEVA; TSIPORKOVA, 2014; LI, C. et al., 2015; CHEN, X.; ZHOU, M.; CARIN, 2012; ZHENG et al., 2011) but might include documents' meta-data, such as publication-venue and co-authorship information (THO; HUI; FONG, 2003; J et al., 2016). Additional information might also be added, such as work relationships (BALOG; RIJKE, 2007). Other work cluster experts based on their expertise representation, using techniques such as structural regularity (VAN GYSEL; RIJKE; KANOULAS, 2017).

While most studies concentrate on content-based topic model similarity (LI, C. et al., 2015; CHEN, X.; ZHOU, M.; CARIN, 2012; ZHENG et al., 2011), others examine similar authors cited together (THO; HUI; FONG, 2003) or have social proximity based on previous collaboration information (J et al., 2016).

Boeva et al. (BOEVA; BONEVA; TSIPORKOVA, 2014) introduce an expert clustering approach by partitioning experts based on the keywords associated with their documents. To extract these keywords from the expert's documents, they apply a part-of-speech tagger to the documents' data and extract three types of keywords: (i) *adjective-nouns* - an adjective followed by a noun; (ii) *multiple nouns* - sequence of nouns; and (iii) *single noun* - the remaining nouns. Once all experts' keywords are extracted, they are clustered through a semantic similarity metric based through Wordnet. Each expert profile is transformed into a vector, where each dimension represents the percentage of keywords in the expert profile present in the given cluster. Once the experts' vectors are built, the Euclidean distance is applied to cluster them and identify similar experts.

2.5.5 Expert recommendation

Expert recommendation (YANG, C. et al., 2015; SUN et al., 2011; COHEN; EBEL, 2013; TANG, Jie et al., 2012; YANG, Chen et al., 2014; CHAIWANAROM; LURSINSAP, 2015; YANG, C. et al., 2015; XU, Y. et al., 2012; CHEN, H.-H. et al., 2011; CABANAC, 2011; KONG et al., 2016; GOLLAPALLI; MITRA; GILES, 2012) (also called *matching* in the literature (YANG, Chen et al., 2014)) 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 experts that could make a worthwhile collaboration through their complementary expertise. For example, a text-sequence processing expert could collaborate with a DNA mapping expert. A fruitful topic in expert recommendation is: how should one match experts from different domains of knowledge, such as medicine and computing? Few studies have addressed this issue (TANG, Jie et al., 2012; ARAKI et al., 2017).

The techniques adopted in the literature to carry out expertise recommendation vary. Most use expertise evidence combined with social relations (COHEN; EBEL, 2013; TANG, Jie et al., 2012; YANG, Chen et al., 2014; CHAIWANAROM; LURSINSAP, 2015; YANG, C. et al., 2015; XU, Y. et al., 2012; CHEN, H.-H. et al., 2011; CABANAC, 2011; ABITEBOUL; BUNEMAN; SUCIU, 2000; ZHOU, X. et al., 2017). Some studies adopt alternative approaches, such as:

- creating a grant database associated with content similarity and the collaboration network to suggest new collaborations between researchers with different areas of expertise (ARAKI et al., 2017);
- Introducing path optimization to graphs linking authors and articles contents, where, for example, a path *Author1-Paper1-Term-Paper2-Author2* becomes an *Author-Term-Author2* path, thus a) resulting in a smaller graph, b) improving random-walk algorithm application (ZHOU, X. et al., 2017);
- Use concepts from the *expertise seeking* area (HOFMANN et al., 2010).

Cohen et al.(COHEN; EBEL, 2013) introduces a researcher collaboration suggestion based on the researcher's social network (collaborations) and a given topic of interest for collaboration, defined by keywords. A graph is built including the authors (vertices where they are represented by a bag of words from their publication titles) and collaborations between authors (edges composed of three features: the publication title, date, and venue). Over the graph, a query composed of an author (a vertex) and a set of keywords 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 it is that a given expert will work on the topic specified by the keywords from the query. Two approaches are introduced. The first approach uses TF-IDF between the expert profile and query keywords, while the second uses a self-developed function called *Collab*. *Collab* considers the previous collaborations of a given expert to determine if he/she is relevant to the query. For each previous collaboration, it calculates how relevant it is to the query (TF-IDF on the keywords and collaboration title), how much time has passed since the collaboration (logarithm function), and if it occurred on a venue where the query expert has already published. In its last step, *Collab* sums the previous values from all the neighbors of a given node to calculate the weight (relevance) of the node to the original query node.

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

2.6 CURRENT WORK COMPARISON

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 based on two criteria: first, to cover examples in the full range of taxonomic classifications; and second, to include the most recent or relevant ones. We selected 26 works, which are compared in Tables 2 through 6. 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 surveyed:

- most works use public unrestricted plain-text sources, such as public sites and other data sources with no clearly defined structure. Since the most common expertise extraction methods are based on terms (such as topic models, language models, and term frequency), structured data is not required;
- a complete data extraction process is more common than a focused data extraction. Many works form an expert representation in advance and thus allow browsing and searching for the extracted expertise information. Concerning 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 standard, especially for complex expert composition. It is natural since it is a good way of designing relations such as those between author-venue and author-coauthor, as well as finding citations on documents or people;
- most of the studies focus on expert finding, followed by those focused on expert recommendation. That indicates how finding experts on a given topic is a significant factor in expertise retrieval.

2.7 OPEN ISSUES

This section introduces some open issues on automated expertise retrieval. It is organized by the affected element: the expert, the expertise evidence, or the end user of an expertise retrieval system. Based on this analysis, we describe how our proposal, the Experion framework, may improve or provide support to improve these issues.

2.7.1 Expert-related open issues

Two open issues were identified related to the expert: *expertise association* and *cross-domain collaboration*. Regarding expertise association, automated expertise retrieval must address specific issues:

Table 2 – Current work comparison - 1/5

Work	Data source	Data extraction	Expertise extraction	Task (Application)
Li2015(LI, C. et al., 2015)				
	Private unstructured plaintext	General simple expert composition	Custom temporal continuous with semantic support	Expert profiling (Builds topic cloud views of expertise models based on CVs)
Xu2012(XU, Y. et al., 2012)				
	Public unrestricted unstructured plaintext	General complex expert composition	Graph and custom, with lexical database support	Expert recommendation (Suggests collaborators based on scientific publications)
Chen2013(CHEN, H.-H. et al., 2013)				
	Public unrestricted semi-structured dataset	General simple expert composition	Term frequency extraction, encyclopedia semantic support	Expert finding (Finds experts based on CiteSeer and Wikipedia data)
Fang2014(FANG; GODAVARTHY, 2014)				
	Public unrestricted unstructured plaintext	General complex expert composition, with stemming and stop words removal	Language model extraction, time slices temporal support	Expert profiling (Based on previous publications, analyzes if an author may change its research line in the future)
Parada2013(PARADA et al., 2013)				
	Private semi-structured and structured dataset	General complex expert composition	Graph extraction	Expert recommendation (Uses a social-based calculated PCI - Potencial Collaboration Index - to recommend collaborations)
Pal2015(PAL, 2015)				
	Public unrestricted unstructured plaintext	Focused simple expert composition	Topic model extraction	Expert finding (Combines multiple data sources to locate an expert)

Table 3 – Current work comparison - 2/5

Work	Data source	Data extraction	Expertise extraction	Task (Application)
VanGysel2016(VAN GYSEL; RIJKE; WORRING, 2016)				
	Public unrestricted unstructured plaintext	General simple expert composition with stop words	Term frequency extraction	Expert finding (Promotes expert finding optimization using back-propagation neural networks)
Rybak2014(RYBAK; BALOG; NØRVÅG, 2014)				
	Public restricted semi-structured 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 authors changes over time)
Liu2013(LIU, D. et al., 2013)				
	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(GOLLAPALLI; MITRA; GILES, 2013)				
	Public unrestricted semi-structured dataset and plaintext	Focused simple expert composition	Graph and topic model extraction	Expert ranking (Gathers Arnetminer data, applying a modified PageRank and a tripartite graph algorithm)
Boeva2014(BOEVA; BONEVA; TSIPORKOVA, 2014)				
	Public unrestricted unstructured plaintext	General simple expert composition, with stemming and stop words removal	Topic model extraction	Expert clustering (through the authors' profiles keywords, compared using Wordnet)
Kaya2014(KAYA; ALHAJJ, 2014)				
	Public unrestricted semi-structured dataset	General complex expert composition	Term frequency extraction, with time slices temporal support	Expert finding and profiling (builds a data cube based on the publication data and applies OLAP methods to locate and profile experts)

Table 4 – Current work comparison - 3/5

Work	Data source	Data extraction	Expertise extraction	Task (Application)
Chaiwanarom2015(CHAIWANAROM; LURSINSAP, 2015)				
	Public restricted semi-structured dataset	General complex expert composition	Graph and topic model extraction, with time slices temporal support	Expert matching (suggests potential collaborations based on social relations, researcher seniority and publications' content similarity)
Fang2010(FANG; SI; MATHUR, 2010)				
	Public unrestricted unstructured plaintext	Focused simple expert composition, with stemming	Custom and language model extraction	Expert finding (introduces a discriminative model to associate authors to documents)
Ganesh2016(J et al., 2016)				
	Public unrestricted semi-structured and unstructured dataset and plaintext	General simple expert composition	Custom extraction	Expert profiling (uses a neural network to learn how to associate authors to documents)
Mangaravite2016a(MANGARAVITE; SANTOS, R. L., 2016)				
	Public unrestricted semi-structured 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(YANG, C. et al., 2015)				
	Private semi-structured and unstructured dataset, plaintext	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)

Table 5 – Current work comparison - 4/5

Work	Data source	Data extraction	Expertise extraction	Task (Application)
Liu2014a(LIU, X. et al., 2014)				
	Public unrestricted unstructured plaintext	Focused complex expert composition	Language model extraction	Expert finding (Introduces AMiner-mini, a version of Arnetminer applicable in institutions)
Neshati2014(HASHEMI; NESHATI; BEIGY, 2013)				
	Public unrestricted unstructured plaintext	Focused simple expert composition	Language model extraction	Expert finding (locates leading authors in a publication)
Balog2009(BALOG; AZZOPARDI; RIJKE, 2009)				
	Public unrestricted unstructured plaintext	General and focused expert composition, with stop words removal	Language model extraction	Expert finding (Introduces a language modeling framework for expert finding)
Cohen2013(COHEN; EBEL, 2013)				
	Public unrestricted semi-structured dataset	General complex expert composition	Graph and term frequency extraction with time slices temporal support	Expert recommendation (uses a start researchers, keywords and co-authoring network to suggest collaborations)
Zhu2014(ZHU, H. et al., 2014)				
	Public unrestricted unstructured plaintext	General simple expert composition	Topic model extraction	Expert finding (besides experts in the desired area, also includes experts on other related areas)
Deng2012(DENG et al., 2012)				
	Public unrestricted semi-structured dataset	Focused complex expert composition	Graph extraction	Expert ranking (using co-authoring network and citations)
Li2015(LI, C. et al., 2015)				
	Public unrestricted unstructured plaintext	General simple expert composition	Topic model extraction	Expert clustering (analyzes publications' content similarity)

Table 6 – Current work comparison - 5/5

Work	Data source	Data extraction	Expertise extraction	Task (Application)
Kumar2016(KUMAR, V.; PEDANEKAR, 2016)				
	Public unrestricted unstructured plaintext	General complex expert composition	Term frequency extraction	Expert finding (through Q&A site data, considering *best-answers* indications as expertise hints)
Osborne2013(OSBORNE; MOTTA; MULHOLLAND, 2013)				
	Public unrestricted semi-structured and unstructured dataset and plaintext	General complex expert composition	Custom extraction	Expert finding and profiling (introduces the Rexplore tool, to visualize and relate authors publications and expertise)
Sateli2017(SATELI et al., 2017)				
	Public unrestricted unstructured plaintext	General complex expert composition	Graph extraction	Expert profiling

- How can one associate a person with a document?
- If a document is associated with more than one person, who is responsible for each item of expertise evidence?
- How essential or reliable is a document as a means of representing the expertise?

These are not trivial issues (BALOG et al., 2012). Concerning the task of associating people with evidence, the studies in the literature vary a good deal. Some use meta-data (such as bibliographic networks, post authors in social networks/forums, or e-mail header information) (BERENDSEN et al., 2013). Others use the person's name or e-mail address in the document - this can cause problems such as ambiguity in the name or the cited name/e-mail may not indicate actual authorship but be just a reference (HASHEMI; NESHATI; BEIGY, 2013). Finding reliable ways to associate people with evidence is still an open research topic.

In a multiple-author document, the issue of determining how well each author's 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 (LUONG et al., 2015). However, this assumes some semantics, such as the first author would always be the most knowledgeable, which we can not assume is always

true.

The methods employed to establish how essential or reliable the evidence in a given document is to estimate a person's expertise vary by the document type. The number of referral links on web pages can measure reliability (*Page Rank*) (ZHU, Jianhan et al., 2010). In the case of scientific articles, the number of citations can be used (CHEN, H.-H. et al., 2013). Depending on their importance, some studies analyze the author's publications over time and seek to identify the preferred topics(RYBAK; BALOG; NØRVÅG, 2014).

Another topic that has not received attention is expertise collaboration between different domains - *cross-domain collaboration* - for example, between biology and computer science. Cross-domain collaboration is more complex than intra-domain collaboration since identifying the related work between different domains is not a trivial task. For example, computer science research on string similarity and sub-string analysis can be applied to DNA sequencing - but how can we find such a relationship? Some approaches draw on studies in the literature (TANG, Jie et al., 2012) to find possible collaborations. A suggestion would be to use a semantic mediator to identify conceptual relationships and find possible new forms of collaboration. Wikipedia is one example of a possible semantic mediator. A system can be devised where a person inputs his problem description, and related techniques and research could be suggested.

2.7.2 Expertise evidence related open issues

Four open issues were identified related to expertise evidence: *combining multiple pieces of evidence, working with multiple languages, assessing the veracity of data, contextual analysis*, and *implementation/information exchange*. A wide range of features is considered when analyzing expertise evidence, such as topic models, social relations, and semantic analysis through Wikipedia articles. However, few works(CUMMINS; LALMAS; O'RIORDAN, 2010; FANG; SI; MATHUR, 2010) address *combining multiple pieces of evidence* to improve the results. Learning approaches, such as neural networks based on user feedback, could provide new and valuable 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 on a social network should weigh lower than papers published at a prestigious conference.

To the best of our knowledge, none of the existing studies consider that the expertise evidence can be in more than one language. In our view, it is advantageous to correlate the same expertise described in *multiple languages* in several cases. One researcher is attempting to relate knowledge from the same domain in different languages (B et al., 2011), but it is still in its early stages. We suggest using a common semantic mediator, such as Wikipedia, for this task. Wikipedia has several articles and has been

successfully used (CHEN, H.-H. et al., 2013; DAVOODI; KIANMEHR; AFSHARCHI, 2013) 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.

Another open issue that should be pointed out is how to assess the quality/trust of expertise evidence - *data veracity*. Analyzing data veracity is a common problem in big data integration(DONG; SRIVASTAVA, 2013) but has not been studied yet in expertise retrieval. Some features that are found in expertise retrieval that could be combined with an analysis of data veracity include:

- 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;
- An article published in a conference, 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; and
- As measured by established metrics such as the H-Index (HIRSCH, 2005) or JCR (ANALYTICS, 2017), the impact level of scientific publications is also a strong indicator of quality/trust.

Research and expertise evolve as a result of events in the context of the involved people or topic. Thus a *contextual analysis* is essential. For example, researchers in academia may start working on a new topic based on a ground-breaking article (WANG, X.; ZHAI; ROTH, 2013). Professionals in the industry may change their expertise interests as a result of a significant recent event. Awareness of context when analyzing may yield interesting results and assist in understanding the evolution of expertise and tracking changes in the topics of interest over time.

Implementation and information exchange is other open issue. Few studies (KOU et al., 2015; VAN GYSEL; RIJKE; WORRING, 2016) have analyzed questions related to implementation or scalability when introducing their schemes. Thus, issues such as how expertise representation should be indexed and its searches facilitated are open to further suggestions and improvements. Another interesting research area is a standard expertise representation that could be exchanged between systems. In big data, expertise information can be retrieved by several systems. A standard expertise representation could provide a useful way to scale these systems, allowing them to exchange information. It could be used to compare different approaches, finding the best for a given domain.

2.7.3 User-related open issues

Three open issues related to the user of an expertise retrieval system were identified: user interaction, explanation of the results, and description of the expertise. Many methods can be employed for automated expertise retrieval, but to the best of our knowledge, none completely account for user feedback during the process, i.e., *user interaction* (BALOG et al., 2012). Some do so in a limited way, for example, by classifying documents(HIEMSTRA, 2001) or expert matching(TANG, Jie et al., 2010) as relevant, irrelevant, or false.

When including user feedback, a suggestion would be to define the intention of the user when working with expertise retrieval. Some metrics could be used to measure his/her degree of satisfaction with the results while ensuring minimal user interaction. User interaction can be regarded not only as a way to adjust parameters but also to make alterations in design decisions on automated expertise retrieval and related tasks. Thus, this could ensure a more general, interchangeable, and component-based approach quickly adapted to new domains and data formats based on user feedback.

Helping the user understand the results is another topic we found that is ignored in the literature and classified as the *explanation of the results*. The better the user understands the results, the more confidence he/she will have in the expertise retrieval system. A system could be adopted to help the user understand this by, for example, describing how the given expertise was captured and how the system assesses its relevance. The current approaches fail to describe how they obtained such data. Adopting an approach where the user does not understand how the result was achieved is inappropriate 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, he/she is free to use his/her judgment and decide whether or not to go ahead with contacting the referred expert.

The last user-related open issue is directly related to the previous open issue: *description of the expertise*. An expertise description should be clear, concise, and preferably human-readable. To the best of our knowledge, there are few approaches to automatically making a human-readable representation of expertise (BALOG; DE RIJKE, 2007; LATIF; AFZAL; TOCHTERMANN, 2010). This representation could assess a system's quality by comparing it with the expertise obtained from several systems. Naturally, a standard representation of expertise would be required to make a comparison.

2.7.4 Open-issues support and solutions

In the next chapter, we introduce our proposed framework Experion, detailing its concepts and structure. It introduces key concepts (Fact, Dimension and Context) to

provide support to solving three open issues presented in this chapter:

1. **contexual analysis** - we introduce a model to contextualize the expertise using the concept of *Contextual Dimensions* and *Context*;
2. **explanation of results** - using the concepts introduced by Experion a user can have a better understanding of how a person's expertise was obtained;
3. **description of expertise** - existing expert finding systems can describe the expertise of a person by applying our proposed Experion framework to their data.

3 EXPERION

In this chapter, we introduce Experion, our proposal for Expertise representation and contextualization (GONÇALVES; DORNELES, C. F., 2022). Based on collected data from the pieces of evidence, we elaborate on the expertise representation and context. *Experion* is a *black-box framework* that gives the user more information to understand the expertise while reusing/improving existing work in expert finding. A framework is a reusable, semi-complete application (FAYAD; SCHMIDT, 1997). Experion implements the basic structure necessary for expertise contextualization in expert finding systems. It adopts a black-box where the framework defines a contract (MEYER, 2007), establishing the expected input and output of the components plugged into the framework.

Experion proposes a standard for (i) representing expertise evidence, (ii) generating expertise description, and; (iii) associating context with the expertise. Our objective in introducing contextualization is to tackle the issue of helping the users understand the results of expert finding.

The better the users understand the results, the more confidence they will have in the process. Current approaches in expert finding usually present a list of people or a graph cluster as a result but do not describe how they obtain such data (CHEN, H.-H. et al., 2013; PAL, 2015). Adopting an approach where the user does not understand how the result was elaborated is inappropriate when dealing with people. Collaborations based on false assumptions could, for example, result in unsuccessful interactions and should be avoided. By allowing the users to understand the result, they can use their judgment and decide whether to go ahead by contacting the referred expert.

An approach to help the users understand the results of an expert finding system is to provide contextualization in the expert finding process results. A context, in this case, should contain information that allows a user to understand how the supposed expertise was obtained or applied. We can cite as examples of contexts: (i) Where the expertise evidence occurred or was demonstrated (institution, venue); (ii) What kind of expertise evidence is it (teaching, research)?; (iii) If it is a solo or group activity and; (iv) The impact of the activity (such as the Impact Factor (IF) in academic publications).

It is important to note that elaborating the context in an automated way is not a trivial task and faces some challenges:

- There may be a need to analyze the expertise evidence as a set by gathering *hints* (such as the same person appearing several times) from other evidence associated with the same year or location, for example;
- External data sources can provide additional data - for example, the Impact Factor (IF) associated with a venue where an article was published; and
- Describing the context in a textual form, concise and natural, is another non-

trivial issue.

Section 3.1 introduces our framework proposal and basic structure. A detailed description of its concepts is introduced in Section 3.2, with an example of how Expertise Contextualization works presented in Section 3.3. Section 3.5 compares our proposal with existing work and Section 3.6 discusses how Experion can improve existing open issues in Expertise Retrieval (detailed in Chapter 2).

3.1 EXPERION OVERVIEW

Following the concept of a *black-box framework* based on a *contract interface*, **Experion** components can be integrated into a traditional expert finding system to improve the results with contextualization. A traditional expert finding system is composed of a search process, where given an expertise of interest, it locates evidence of such expertise and ranks the candidate experts associated with it. Experion introduces new components in this process, expanding it and focusing on improving the result presentation and contextualization. Figure 3 describes an expert finding process with the framework applied to it.

First, Experion introduces a standard representation of expertise evidence (called *Fact*) and their associated data (called *Dimensions*). A Dimension that contains context data is called a *Contextual Dimension*. An expert finding system can use these representations to generate a list of Facts (and their associated Dimensions) associated with a candidate expert. These Facts can be generated previously, *offline*, stored in a database, or generated on the fly during the search process. Since Experion focuses on providing the representation and the *contracts* of the process, it is up to the expert finding system to define the best moment and how to generate the Facts and their associated Dimensions. Given that Experion provides a standard representation, common function libraries could be shared between expert finding systems to implement this generation. How to construct such libraries is outside the scope of this work and is considered future work. Steps **1 and 2** in Figure 3 contemplate this process.

Steps **3 through 6** in Figure 3 are the new steps in an expert finding provided by Experion. First, it introduces the *Derivators*, functions that, given a set of *Facts* and their associated Dimensions, can generate a special kind of Dimension, called *Derived Dimension*, that expands and standardizes the contextual information. The framework does not provide a pre-defined set of Derivators but defines the contract to them - receive the Facts and associated Dimensions and generate Derived Dimensions associated with these Facts. Thus, a Derivator can be used in several expert finding systems - a public, shared library could be constructed. Steps **3 and 4** in Figure 3 contemplate this process.

Up to this moment, the expert finding process can produce, associated with the candidate experts it found, a list of expertise evidence (Facts) with contextual data

(Derived Dimensions) associated with them. To further improve on the contextualization, another concept is introduced by Experion - the *Context Builder*. Context Builder functions, given a set of Facts and their associated Derived Dimensions, can create a *Context*, i.e., a description of the context associated with a Fact. Step **5** introduces this process. A Context can be as simple as just a list of the Derived Dimensions or as complex as a natural language description built using the Derived Dimensions. As in the case of the Derivators, the framework provides a contract for a Context Builder, which receives as input a list of Fact and Derived Dimensions and produces contexts, which can be associated with one or more Facts. Public shared libraries of standard Derivators could be built and are contemplated in future work.

With the built Contexts, the expert finding system can provide a contextualized result to the user. It can show which expertise evidence (Facts) is used to consider a given person a candidate expert and the Context associated with this expertise evidence. Step **6** in Figure 3 introduces this last step.

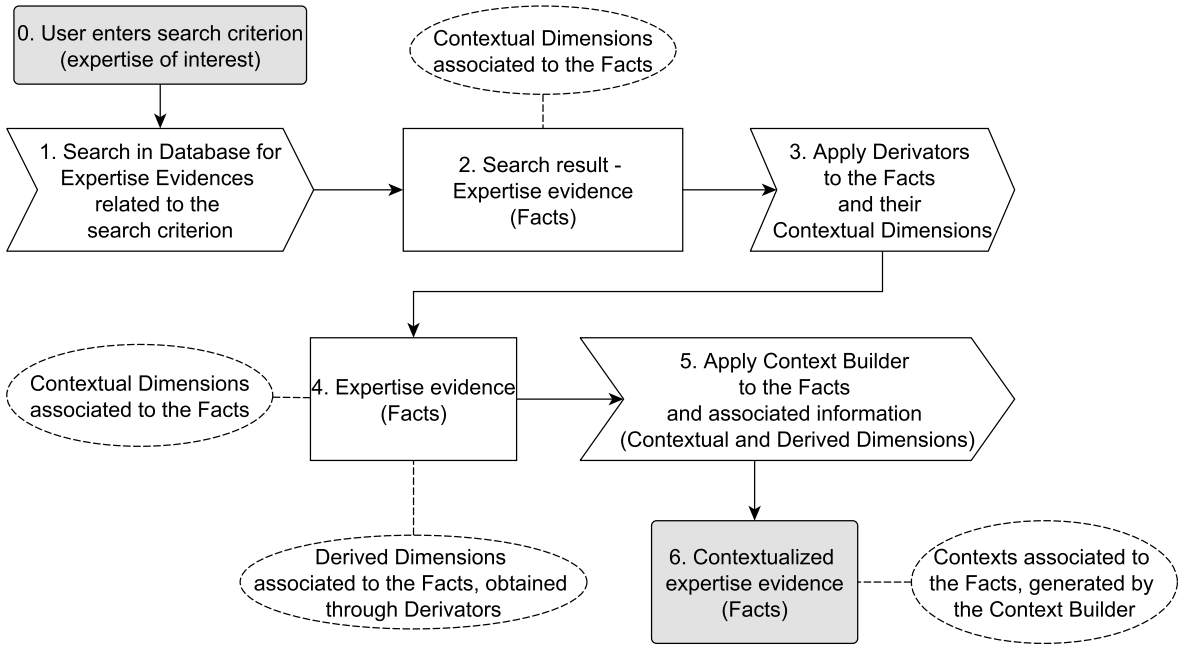


Figure 3 – Experion overview

3.2 CONCEPTS

This section details the key concepts over which the Experion framework is structured: Entity, Fact, Context, Derived Dimensions, Derivators, and Context Builder.

3.2.1 Entities

An **Entity** is the final object of interest for the expert retrieval process. It can be a person (expert), an institution (university) or a entity (research group). Instead of expert

(as in existing work), the Entity term was chosen to generalize the framework, allowing it to be applied to processes with alternative outputs - such as finding an Institution of interest instead of a single expert. An expert finding system will output a ranked list of Entities.

Every other concept in the framework is connected somehow to an Entity. An entity is composed at least by a unique property, named *id*, which uniquely identifies it. The framework does not limit which additional fields may be present. Thus, one could specialize an Entity to specific classes, such as an expert, which could have a name and a date of birth. It could also be specialized to a University, which would have a name and a website, for example. Figure 4 demonstrates such specialization. As introduced by the framework, we have the Entity concept specialized in two kinds of Entity, Expert and University. Expert introduces DOB - Date Of Birth and Name. University introduces Name and URL.

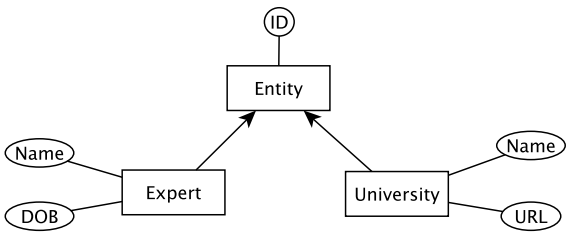


Figure 4 – Entity Concept

3.2.2 Fact

A Fact is the framework representation of Expertise Evidence. During the expertise retrieval, each piece of evidence of expertise found for a given Expert is mapped to a Fact. Per itself, a Fact contains three properties: (i) an *id*, which uniquely identifies it; (ii) a start year; (iii) and an end year. We defined a start and end year instead of simply a year because some evidence occurs in several years. For example, a piece of evidence representing a Master's Degree course can span two or more years. For single-year evidence (such as an article publication), the start and end year contain the same value. For an ongoing Fact, the current year can be assumed as the last year. We chose *year* as a granularity to represent time (instead of a month or even a date) due to the fact that we're dealing with expertise. Normally, a person won't acquire/focus on a new expertise quickly enough that a granularity smaller than year would be necessary.

Similar to the Entity, a Fact could be specialized to represent particular situations. For example, one could have a HearsayFact, which would represent a Fact based on something said that has low confidence. That could be, for example, a testimony about a researcher given by a colleague. Such specialization could be used to treat such facts differently when elaborating on the expertise. Figure 5 demonstrates this specialization.

Although we suggest such specialization, the same could be achieved directly using a specific Dimension associated with such HearsayFacts. The following Section introduces how one could do it and how the dimensions relate to a fact.

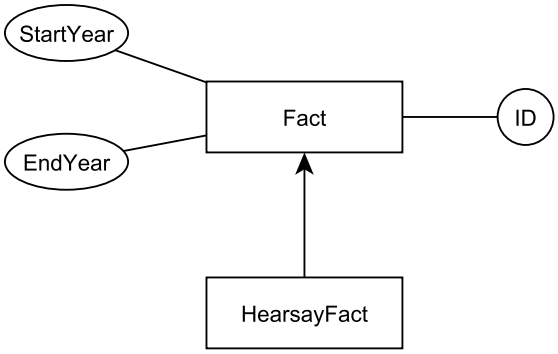


Figure 5 – Fact Concept specialization

3.2.3 Dimension

A *Dimension* is any information contained or related to a Fact. A Dimension is not exclusive to a single Fact - it can be shared between two or more Facts. Such structure allows for identifying related Facts based on their Dimensions. A Dimension implementation must have two primary actions: (i) it can compare itself to another Dimension, indicating if they mean the same thing, and; (ii) it can be combined in a single Dimension. If two Dimensions are equivalent, the action or merging can be requested, which produces a single Dimension from the original Dimensions. For example, a Collaboration Dimension, which lists the people who collaborated on a given Fact, can be merged by combining the list of collaborators' names in the two Collaboration Dimensions. The framework establishes two specializations for a Dimension: IndexDimension and ContextualDimension. An IndexDimension is a piece of information that can be used for locating Facts given search criteria but does not represent a context associated with the Fact. For example, the abstract of an Article can be used as an IndexDimension but is not applicable to contextualize such an Article. A ContextualDimension is a piece of information that can be used to contextualize a Fact—for example, the venue of an Article.

Besides these two specializations, the framework also established a third specialization, for the ContextualDimension, called DerivedDimension. A DerivedDimension is a special kind of Dimension that is not extracted directly from the source data but is generated dynamically throw the contextualization process in the framework. Such a task is the responsibility of the framework Derivators, introduced in the following Section. Figure 6 demonstrates the Dimensions kinds. As an example of an IndexDimension, we have a *Keywords* dimension that can contain the keywords associated

with an article. For ContextualDimensions, we have two examples: *Institution* (for an Examining Board) and *Venue* (for an article). Lastly, we have one DerivedDimension called *Location*, which could be generated as a standardization of the Institution and Venue ContextualDimensions.

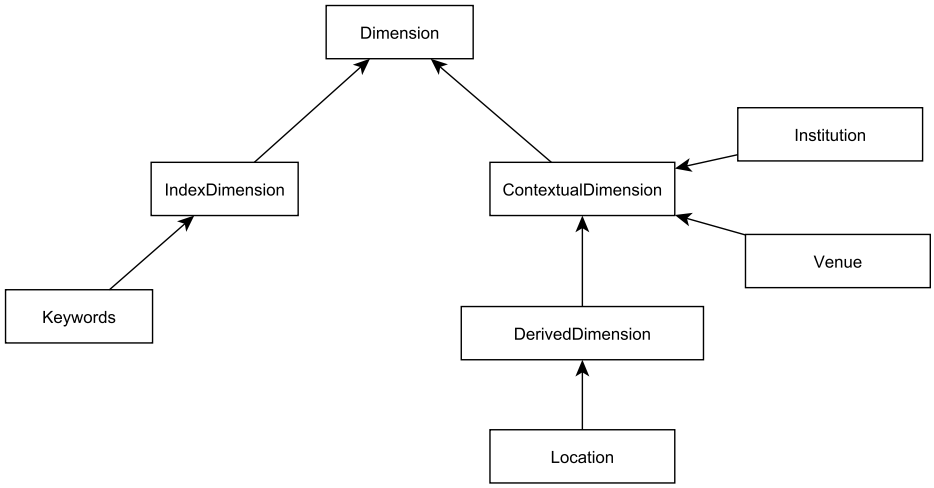


Figure 6 – Dimension Concept and its specializations

3.2.4 Derivators

The Experion framework introduces the ContextualDimension concept to provide contextualization. However, although such Dimensions provide contextualization, they may not be normalized - i.e., we could have different ContextualDimension that, in the end, represent the same context. For example, we could have two ContextualDimension where one represents studying in a Technical School, and another represents studying in a University. From both ContextualDimensions, we can identify a context of studying beyond regular school. However, based on the two ContextualDimension alone, the framework would not be able to relate the facts associated with such dimensions.

The DerivedDimension concept was established to normalize the context associated with the Facts. A DerivedDimension is generated during the expertise retrieval and contextualization process based on the existing ContextualDimensions. In the framework context, such a process is a responsibility of a *Derivator Function*, or *Derivator* for short.

A Derivator receives a set of Facts and, based on their ContextualDimension, can produce DerivedDimension. How this process is done varies considerably based on the data available and which kind of context is of interest. In the context of Academia, a Derived Dimension could be, for example, "activity," which can be "teaching" or "research." Another example could be "degree," which could be "undergraduate" or "graduate."

Besides existing data in the Facts, a Derivator can use (but is not obligated to) any external system or data-source (for semantic support, as an example) to produce

its result. We could have a ImpactFactorDerivator that, given a Fact representing a published article, would identify the Venue associated with it and retrieve the Impact Factor of the Venue, adding it as a new ContextualDimension. This could be used later to improve result ranking in an expert finding system.

An example of a Derivator could be a function that, given a Fact associated with teaching (a class lecture, for example) and a university, can derive an activity of "teaching" and a "degree" of "graduate." Thus, from the original Facts and associated ContextualDimensions, new DerivedDimensions activity and degree can be generated. Using the ContextualDimensions and DerivedDimensions the Experion framework provides a basis for Expertise contextualization, allowing the creation of one or more contexts, introduced in the following.

3.2.5 Context

A **Context** is an abstraction that relates a Fact with a set of DerivedDimension, contextualizing the Fact under different perspectives: where it happened, when it happened, how it happened, who is related to it, and so on. In expert finding, a DerivedDimension could be, for example, "activity," which can be "teaching" or "research." Another example could be "degree," which could be "undergraduate" or "graduate." Thus, from the two Derived Dimensions ("teaching" and "graduate"), a Context of "graduate teaching" can be established.

Since a set of DerivedDimensions defines a Context and the DerivedDimensions are shared between Facts, a single context can be associated with several Facts. At the same time, a single Fact can be associated with different Contexts since the framework can generate different contexts by applying different Context Builders, introduced in the next Section. Figure 7 demonstrates the proposed relation between the concepts.

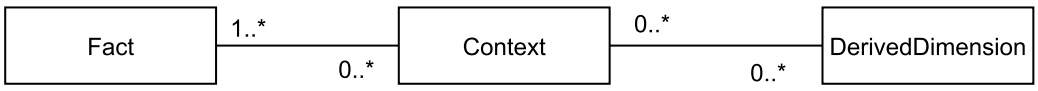


Figure 7 – Context Concept

3.2.6 Context Builder

As introduced, a **Context** is built from ContextualDimension and DerivedDimension 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 Dimensions and can produce one or more Contexts. Chapter 4 introduces an example of a Context Builder we developed, the **Simple Weighted Context Builder** (SWCB).

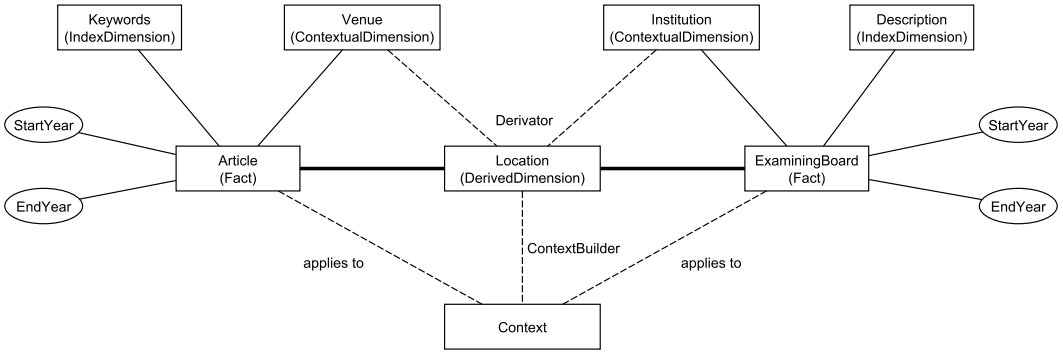


Figure 8 – Concepts application example

Figure 8 provides an example of applying Fact, Dimension, and Context concepts in academia. Two types of Facts are present: **Article** and **ExaminingBoard**. Each of these types has associated Dimensions (naturally, there are others we ignored in this example for simplification purposes). Article has a *Keywords* IndexDimension that contains its associated keywords and a *Venue* ContextualDimension. ExaminingBoard has two associated dimensions: the *Description* of the work examined by the board (IndexDimension) and the *Institution* where the examination occurred (ContextualDimension). By applying a Derivator function, a DerivedDimension called *Location* is built and associated with the Facts. This dimension, in the example, standardizes the Institution and Venue contextual dimensions since both represent the same idea: a Location where a Fact occurred. Given the Facts with the DerivedDimension associated, a ContextBuilder function is applied to them, generating a Context. In our example, this would be, for example, an indication of the Location where they occurred. It should be noted that existing ContextualDimensions are not suppressed by the generated DerivedDimensions. Both kinds of Dimensions are kept in the proposed model, since a DerivedDimensions does not necessarily substitutes the ContextualDimensions based on which it was generated.

3.3 RESULT CONTEXTUALIZATION

Figure 9 details the pipeline of an example implementation of the process of expert finding with contextualization, implemented using the proposed Experion framework and the introduced concepts. 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. Here, 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)**. Derived Dimensions are elaborated as a result of their application **(4)**. As described earlier, these Dimensions are generated using the information in the Fact and analyzing the information from other Facts from the set, allowing a certain level of inference.

The framework's next step is applying a Context Builder over the set of Facts and Derived Dimensions **(5)**. The Context Builder analyzes the Derived Dimensions and establishes one or more contexts **(6)** associated with one or more Facts. For each pair (Fact, Context), it calculates a Confidence Factor (Cf) **(7)**, which indicates how strongly the Fact contributes to the Context definition.

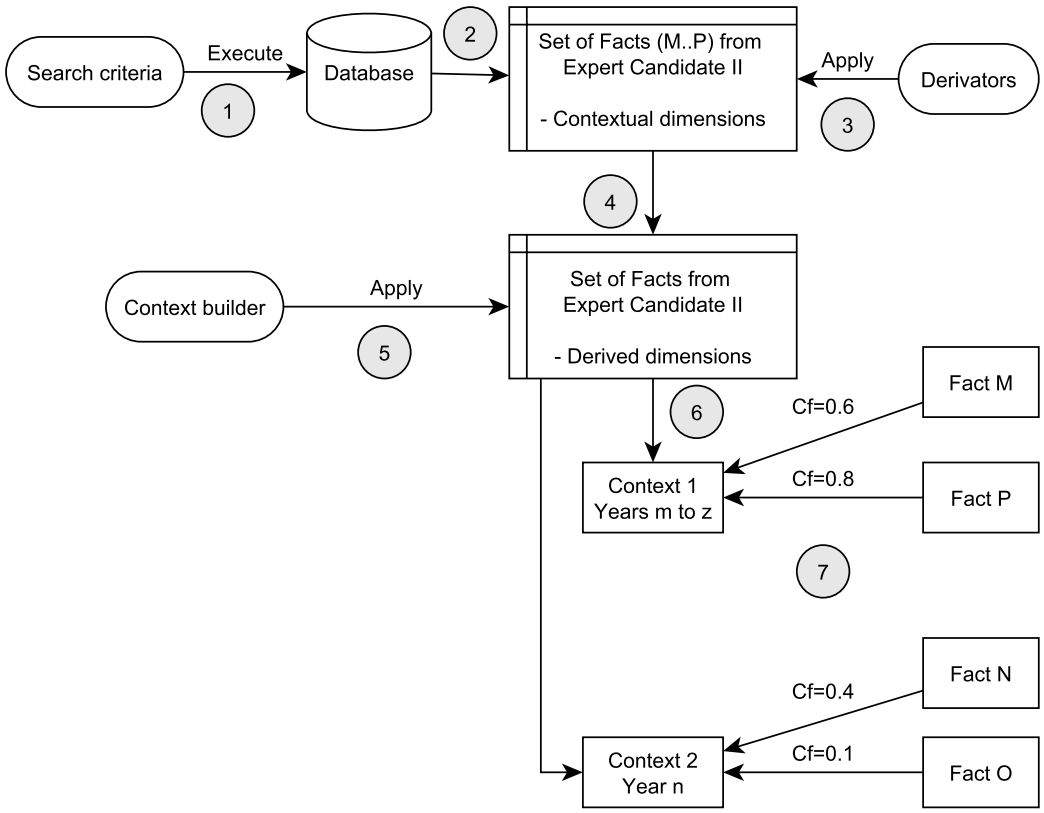


Figure 9 – Contextualized search example

3.4 FRAMEWORK EXTENSIBILITY

Figure 10 provides an overview of all framework concepts and functions, indicating how they interact and in which points Experion is extensible. As introduces earlier, the Fact concept can be specialized to contemplate any kind of expertise evidence - in the example, two specializations are introduced: ExaminationBoard and Article. Similar to Fact, all three kinds of Dimensions (Index, Contextual and Derived) can be specialized as well. In the example there are several Dimension specializations

(Keywords for IndexDimension, Venue for ContextualDimension and ImpactFactor for DerivedDimension).

Besides extensibility by specialization of concepts, Experion also allows developing and integrating new functions to its process, to allow the generation of new kinds of DerivedDimensions (through the Derivator functions) and also alternative Context generations (using ContextBuilder functions). Derivators and ContextBuilder are basically functions that can be applied to Facts and their associated Dimensions to generate DerivedDimensions and Contexts, respectively. In Figure 10 the SWCB function is presented as an implementation of a ContextBuilder and ImpactFactorDerivator as an example of a Derivator function.

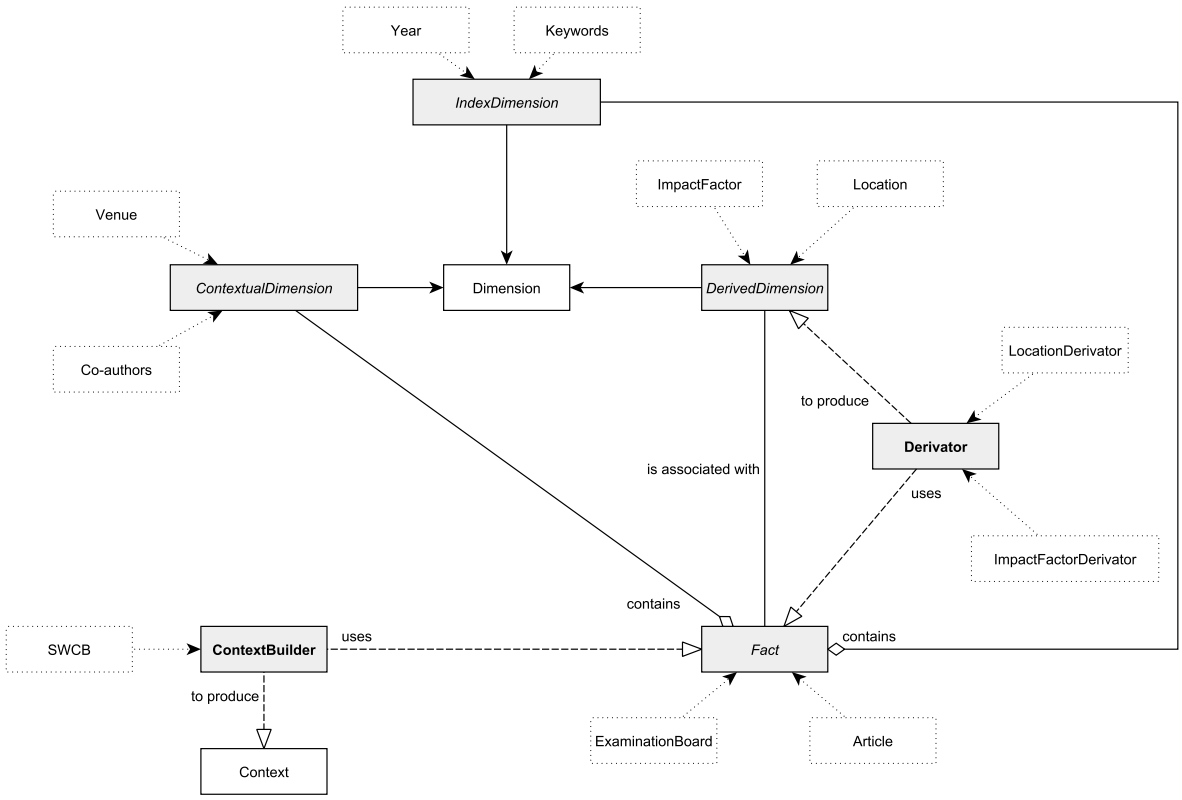


Figure 10 – Experion Framework Extensibility

3.5 COMPARISON TO RELATED WORK

To our knowledge, no current work in expert finding proposes contextualizing expertise as the Experion framework. Existing expert finding works: (i) do not describe the expertise but only list the most favorable candidate experts (with a degree of confidence) or; (ii) describe the expertise in a summarized, algorithm-focused way (such as topic models - i.e., list of words that represent the expertise). Table 7 compares Experion to some related work. Although Experion focuses on expert finding, we include

Work	ScholarLens (SATELI et al., 2017)	PROFILES (M WEBER, 2019)	EOSK (WU et al., 2013)	ArnetMiner (TANG, Jie et al., 2008)	Social profile (BALOG; RIJKE, 2007)	Expertise Manager (LIU, Ping; CURSON; DEW, 2005)	Experion
Expertise data source	Internal	Internal	Internal	Internal	Internal	Multiple	**Existing**
Semantic support	Required	No	No	Required	Required	Required	**Supports**
Temporality	No	No	Yes	Yes	No	No	**Yes**
Context support	No	No	No	No	No	No	**Yes**
Data exchange	Yes	Yes	No	No	No	Yes	**Possible**
Standard representation	Yes	No	Yes	No	No	No	**Yes**
Expertise description	Partial	Yes	No	Yes	No	No	**Yes**

Table 7 – Experion compared to related work

expert profiling-related work in the comparison because Experion's proposal generates a specialized and focused profile as part of an expert finding system result.

We considered seven aspects to compare Experion with existing work. Regarding the expertise data source, Experion does not require specialized data sources and can work on sources already used in expert finding systems. By comparison, related work builds internal data sources specific to their processes. To automatically build the competencies profiles, **ScholarLens** make some assumptions for its work-flow: (i) complete access to the articles' full text; (ii) automatic topic extraction, i.e., relevant named entities identification; and (iii) semantic representation of the extracted data, based on semantic vocabularies. **PROFILES** introduces a platform for registering researchers and their associated publications and activities. It imports and analyzes documents related to a researcher (such as white pages and publications), building an electronic curriculum vitae. **EOSK** builds expert profiles from enterprise microblogs posts and Human Resources information (hierarchy, project participation). The retrieved data is inputted into their proposed ontology. To extract the researcher profiles from data on the web, **Arnetminer** uses a process based on an extended version of the FOAF ontology. Initially, a search is made in Google using the researcher's name to collect pages associated with him/her. A binary classifier is used to identify if the page introduces/describes the researcher based on an SVM learning model elaborated based on previously manually tagged pages. **Social profile** introduces a simplified expert profile, called social profile, based on researcher expertise and previous associations. **Expertise Manager** proposes a brokering system to gather expertise data about a researcher from several sources, using RDF as the common format and integrating this data in a centralized Expertise Manager database.

Regarding semantic support, Experion establishes contracts for its Derivator and ContextBuilder functions. How they operate and which data sources they consider are open, and an implementation could use additional semantic support without changing the framework. Thus, although it can support it, it is not a requirement as in other works. **ScholarLens** requires DBPedia support, **ArnetMiner** uses an extended version of the FOAF ontology, and **Social profile** assumes a list of known knowledge areas based on which the expertise is estimated. **Expertise Manager** requires a conceptual model of the domain to which it is applied.

Temporarility is supported in Experion by providing the concepts of StartYear and EndYear associated with the Facts. Thus, a ContextBuilder can build contexts con-

sidering the associated years. Chapter 4 introduces an example of such ContextBuilder, which considers temporality. Among compared work, only **EOSK** and **ArnetMiner** consider temporality in their expertise analysis.

Context support is the key differential for Experion: no related work considers contextualizing the expertise evidence. Some, such as **ScholarLens**, describe user expertise based on ontology terms and do not include or consider contextualizing the expertise in the terms proposed by Experion (where, with whom, and so on).

Experion allows data exchange given its proposal of a standard representation of Expertise (based on the concepts of Fact and Dimensions). One or more data sources can be parsed, and their information converted into this standard representation (as we show in the implementation of our framework in Chapter 4). Among related work, **ScholarLens** provide more advanced data exchange through ontologies but focuses only on the expertise information and does not consider the contextual information associated with the evidence, as Experion is capable. In **PROFILES** each institution that uses PROFILES has its installation and, since the tool has federation support and is ontology-based, it permits the institutions to share their stored data, building a researcher network. **Expertise Manager** also allows a certain level of data exchange through its broker approach.

Lastly, in Expertise description, we consider if the work is capable of producing an expertise description understandable to a user. Experion, with its concepts and functions provide advances compared to existing work in expert finding. By allowing a contextualized description of the expertise associated with an expert, it improves existing work that only presents the expert and, at most, their associated expertise in a non-standard and human-readable format. **ScholarLens** and **Expertise Manager** provide profiling based on the experts' competencies. **PROFILES** and **Arnetminer** provides an expert profile that focus on the expertise evidence and considers a limited contextualization, only regarding the social relations of the expert.

3.6 OPEN-ISSUES SUPPORT

Among all the open issues listed in Chapter 2, Experion focuses on improving the presentation and understanding of results in expert finding systems. Nonetheless, Experion could be used to at least provide support to deal with other discussed open issues:

- **Expertise association** - Experion model (Entity, Fact, and Dimensions) holds a direct relationship between a person and its associated data information. Thus, based on the data held by Experion, better expertise association techniques can be tried and compared;
- **Combining multiple evidence** - With a standard for representing expertise

evidence, Experion could allow elaborating techniques to compare and combine these evidence;

- **Multiple languages** - The language of a given expertise evidence (Fact) could be an additional Dimension, and an extra step in the workflow of the framework could be implemented to identify and translate expertise evidence as needed;
- **Data veracity** - The source of each expertise evidence could be stored as a Dimension. With this information, a Derivator could analyze it and generate a DerivedDimension reporting the veracity of a given expertise evidence.

4 FRAMEWORK CASE STUDY

In this chapter, we introduce an application of the Experion framework in the context of Academia, more specifically over the Lattes Platform Curricula. The platform allows exporting a researcher's curriculum as an XML file. Figure 11 shows an excerpt of a curriculum obtained in the platform.

We chose the Lattes Platform as our case study due to its semi-structured format (based on XML), with an adequate amount of meta-data associated with the information about the researchers. This allowed a generation of context in a simplified way by extracting the data and applying some basic processes over them (*Derivators*), focusing the case study on the framework structure and the analysis on whether context improves or not the understanding of the results of an expert finding system. Other sources, such as ResearchGate or GoogleScholar, could have also been used as well. To improve the information in the Lattes curricula we used data from the CrossRef[1] database, as shown in Section 4.3.1.

4.1 FACT AND DIMENSIONS EXTRACTION

There are several kinds of Facts (Expertise Evidence) in a Lattes Curriculum. We elected eight kinds to process, which we list in the following. The dimensions underlined were used as IndexDimension, and the remaining were used as ContextualDimension:

- **Award** contains a title, a year, and an institution's name;
- **Education** contains content (description), institution, keywords, year, degree, and a list of tutor names;

[1] https://www.crossref.org/


```
<?xml version="1.0" encoding="ISO-8859-1" standalone="no"?>
<CURRICULO-VITAE SISTEMA-ORIGEM-XML="LATTES_OFFLINE" NUMERO-IDENTIFICADOR="[illegible]" DATA-ATUALIZACAO="04072022" HORA-ATUALIZACAO="114042">
    <DADOS-GERAIS NOME-COMPLETO="Carina Friedrich Dorneles" NOME-EM-CITACOES-BIBLIOGRAFICAS="DORNELES, C. F.;Dorneles, Carina F.;Dorneles, Carina Fried
        <RESUMO-CV TEXTO-RESUMO-CV-RH="Professora no Departamento de Informática e Estatística (INE) da UFSC. Membro do CD-CEBD 2018/2020 ( Comitê Dire
        <OUTRAS-INFORMACOES-RELEVANTES OUTRAS-INFORMACOES-RELEVANTES=""/>
        <ENDERECO FLAG-DE-PREFERENCIA="ENDERECO_INSTITUCIONAL">
            <ENDERECO-PROFISSIONAL CODIGO-INSTITUICAO-EMPRESA="[illegible]9" NOME-INSTITUICAO-EMPRESA="Universidade Federal de Santa Catarina" CODIGO-OR
        </ENDERECO>
        <FORMACAO-ACADEMICA-TITULACAO>
        <ATUACOES-PROFISSIONAIS>
        <PREMIOS-TITULOS>
    </DADOS-GERAIS>
    <PRODUCAO-BIBLIOGRAFICA>
        <TRABALHOS-EM-EVENTOS>
        <ARTIGOS-PUBLICADOS>
            <ARTIGO-PUBLICADO SEQUENCIA-PRODUCAO="315" ORDEM-IMPORTANCIA="">
                <DADOS-BASICOS-DO-ARTIGO NATUREZA="COMPLETO" TITULO-DO-ARTIGO="Approximate data instance matching: a survey" ANO-DO-ARTIGO="2011" PAIS-I
                <DETALHAMENTO-DO-ARTIGO TITULO-DO-PERIODICO-OU-REVISTA="Knowledge and Information Systems" ISSN="02191377" VOLUME="27" FASCICULO="" SER
                <AUTORES NOME-COMPLETO-DO-AUTOR="Carina Friedrich Dorneles" NOME-PARA-CITACAO="Dorneles, Carina Friedrich" ORDEM-DE-AUTORIA="1" NRO-ID-C
                <AUTORES NOME-COMPLETO-DO-AUTOR="Gonçalves, Rodrigo" NOME-PARA-CITACAO="Gonçalves, Rodrigo" ORDEM-DE-AUTORIA="2" NRO-ID-CNPQ="0867460615
                <AUTORES NOME-COMPLETO-DO-AUTOR="Santos Mello, Ronaldo" NOME-PARA-CITACAO="Santos Mello, Ronaldo" ORDEM-DE-AUTORIA="3" NRO-ID-CNPQ=""/>
            </ARTIGO-PUBLICADO>
            <ARTIGO-PUBLICADO SEQUENCIA-PRODUCAO="319" ORDEM-IMPORTANCIA="">
        </ARTIGOS-PUBLICADOS>
        <LIVROS-E-CAPITULOS>
        <TEXTOS-EM-JORNAIS-OU-REVISTAS>
        <DEMAIS-TIPOS-DE-PRODUCAO-BIBLIOGRAFICA>
        <ARTIGOS-ACEITOS-PARA-PUBLICACAO>
    </PRODUCAO-BIBLIOGRAFICA>
    <PRODUCAO-TECNICA>
    <OUTRA-PRODUCAO>
        <ORIENTACOES-CONCLUIDAS>
            <ORIENTACOES-CONCLUIDAS-PARA-MESTRADO SEQUENCIA-PRODUCAO="361">
                <DADOS-BASICOS-DE-ORIENTACOES-CONCLUIDAS-PARA-MESTRADO NATUREZA="Dissertação de mestrado" TIPO="ACADEMICO" TITULO="Segmentação automáti
                <DETALHAMENTO-DE-ORIENTACOES-CONCLUIDAS-PARA-MESTRADO TIPO-DE-ORIENTACAO="ORIENTADOR_PRINCIPAL" NOME-DO-ORIENTADO="[illegible]
            </ORIENTACOES-CONCLUIDAS-PARA-MESTRADO>
        </ORIENTACOES-CONCLUIDAS>
    </OUTRA-PRODUCAO>
    <DADOS-COMPLEMENTARES>
</CURRICULO-VITAE>
```


Figure 11 – Example of a Lattes curriculum

- **Examining Board** contains content (description), an institution, a student, a degree (Undergraduate, graduate, Ph.D.), board members, and a year;
- **Orientation** includes a title, a student name, an institution, a degree, and a year;
- **Published Book** included a title, keywords, co-authors, and a year;
- **Resume** contains content (self-written biography by the researcher).
- **Event Article** contains a title, abstract text, event name, co-author list, and a year;
- **Periodic Article** contains a title, abstract text, periodic name, co-author list, and a year.

We defined five DerivedDimensions to be generated based on the Facts and Dimensions present in the Lattes Curricula. **Collaboration** indicates which people have collaborated with the Researcher (Entity) in the given Expertise Evidence (Fact). It is composed of the names of the people that could be extracted from the Fact ContextualDimension data. The Derivator created to generate such Collaboration DerivedDimension works as follows:

- *Education:* it is the list of tutors' names;
- *Examining Board:* it is the names of the examining board members;
- *Orientation:* the name of the student;
- *Published Book:* the name of the co-authors of the book;
- *Event Article:* the name of the co-authors;
- *Periodic Article:* the name of the co-authors.

The **Cooperation** DerivedDimension indicates that a given Fact is a collaborative work, not an individual result. It is a *flag* so that when it is present, it indicates collaborative work, where more than one person contributed to. Naturally, one could look at the number of people in the Collaboration DerivedDimension and infer what the Cooperation DerivedDimension provides. The intent here is to demonstrate that a system can generate any number of DerivedDimensions that fit its purpose. To detect if a given Fact should have the Cooperation DerivedDimension, the corresponding Derivator considers:

- *Event Article:* if there are other authors besides the Researcher;
- *Periodic article:* if there are other authors besides the Researcher;
- *Book:* if there are other authors besides the Researcher;
- *Examining Board:* if the examining board is composed of other people besides the Researcher.

Level is the academic level associated with a Fact (Undergraduate, Graduate, or Ph.D.). It is generated by its Derivator as follows:

- *Education:* directly from the Education degree ContextualDimension;
- *Examining board:* calculated as the superior level from the degree ContextualDimension. For example, if the degree is Graduate, the derived Degree is assumed to be Ph.D. since it is a common requirement to be part of a Graduate Examining Board;
- *Orientation:* calculated as the superior level from the degree ContextualDimension. It follows the same approach used for the Examining Board;

Location can be a virtual location (such as a periodic/event name) or a physical location (institution name). It is composed of the following information:

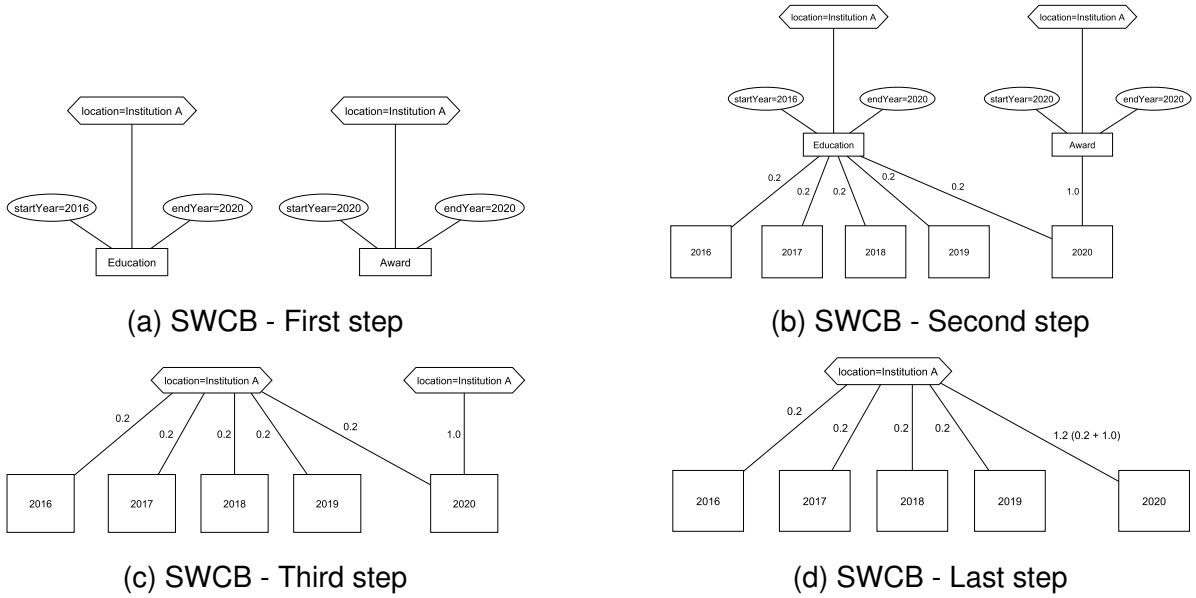
- *Award:* the institution ContextualDimension;
- *Education:* the Institution ContextualDimension;
- *Examining Board:* the Institution ContextualDimension;
- *Orientation:* the Institution ContextualDimension;
- *Event Article:* the event name ContextualDimension;
- *Periodic Article:* the periodic name ContextualDimension;

The last DerivedDimension introduced is **Tutoring**, which indicates that the Fact represents an academic orientation. It is always present for the Orientation Fact, the unique type of Fact representing an Orientation in our data source.

4.2 CONTEXT BUILDING

To build the Contexts associated with the retrieved Facts, we developed the **Simple Weighted Context Builder** (SWCB) as a simple and essentially "proof of concept" Context Builder. It operates as follows:

1. A *Context Bucket* is created for each year associated with the Facts retrieved. If two or more Facts occur in the same year, they share the Context Bucket for that year;
2. Each Context Bucket contains a set of Facts and a set of Derived Dimensions (associated with the Facts in the bucket). Each Fact has a weight associated with the Context Bucket, calculated as the inverse value of the number of years associated with the Fact.
3. Thus, if a Fact occurs in a single year, its weight to that year bucket is 1.0. If it occurs in two years, then the weight for each year is 0.5. This approach is based on the understanding that Facts that occur in fewer years should have a more substantial impact in the context of these years than Facts that span



(a) SWCB - First step

(b) SWCB - Second step

(c) SWCB - Third step

(d) SWCB - Last step

Figure 12 – Simple Weighted Context Builder Example

over several years, since we are contextualizing the environment associated with the expertise. For example, consider a researcher that worked for 20 years in a given university. During this period, in a certain year, he participated in a project in another institution. For that year, his working place on that given institution should have a higher weight on the context of that year than its original working place at his main institution. We chose 1.0 as the base weight for a Fact to normalize the maximum weight a Fact can imply for a context in a single year. This way we treat all Facts as equal on how they impact on the context building.

4. To describe itself, each Context Bucket identifies the equivalent Derived Dimensions and combines them. Each remaining DerivedDimension's associated weight is calculated as the sum of their associated Facts in the bucket where they are located. Thus, if a Derived Dimension is associated with two Facts (by merging their original Derived Dimensions) and the Facts have weights of 0.5 and 1.0, the Derived Dimension weight would be 1.5.

To exemplify the SWCB, let us consider an example. Initially (Figure 12a), we have two Facts: one is an Award received in 2020. Another is an Education related to a Ph.D. between 2016 and 2020. Both have a Location DerivedDimension with the value "Institution A." The SWCB would create four Context Buckets, one for each year from 2016 to 2020. Next, it would associate the Award Fact to the 2020 bucket with a weight of 1.0 and then associate the Education Fact to all buckets with a weight of 0.25 (1.0 divided by four years). This process is shown in Figure 12b. Next, it considers only the

Fact	IndexDimensions	ContextualDimensions	DerivedDimensions	StartYear	EndYear
Award	Title	Institution	Location	Year of award	Year of award
Education	Content Keywords	Institution Degree Tutor Names	Collaboration Level Location	Start of course	End of course or current year if in course
Examining Board	Content	Institution Student Name Degree Board Members	Collaboration Cooperation Level Location	Year of examination	Year of examination
Orientation	Title Keywords	Institution Student Name Degree	Collaboration Level Location	Start of orientation	End of orientation or current year if in course
Published Book	Title Keywords	Co-authors	Collaboration Cooperation	Year of publication	Year of publication
Resume	Content			Current year	Current year
Event Article	Title Abstract	Event name Co-authors	Collaboration Cooperation Location	Year of publication	Year of publication
Periodic Article	Title Abstract	Periodic name Co-authors	Collaboration Cooperation Location	Year of publication	Year of publication

Table 8 – Experion concepts applied to Lattes curricula

Dimensions associated with the Facts since they will provide the data to establish the contexts (Figure 12c). Lastly, it would verify that the two Location Derived Dimensions in 2020 represent the same Location and can be merged. Thus, the Location Derived Dimension in 2020 would weigh 1.20, while the same Dimension in other years would weigh 0.2. That means that in 2020 there is a stronger indication of relation (context) to the Location "Institution A" than in other years, as shown in Figure 12d.

Table 8 provides an overview of (i) all Facts considered from the Lattes Curricula; (ii) which data from the Facts was considered and how it was mapped - Index vs. ContextualDimensions; (iii) which DerivedDimensions are obtained from the ContextualDimensions and; (iv) how the temporality is treated for the Facts considered. All DerivedDimensions are calculated dinamically by the framework using the Derivators defined.

4.3 EXPERIMENTS

In order to validate our hypothesis that adding context improves the understanding by a user of an expert finding system, we promoted an experiment where users could try our proposal and comment on whether it improved or not their understanding of the results. This section describes this experiment, including the dataset used, the tool developed, and the results obtained (user feedback). The experiment was based on our proposed application of the framework over the Lattes Curricula Platform.

4.3.1 Dataset preparation

Figure 13 describes the dataset preparation process we executed to generate the dataset for our experiments. We have used data from the Lattes Platform and extracted the curricula from professors at the Federal University of Santa Catarina to 6.481 curricula from several areas of knowledge, including but not limited to Computer Science - this allowed a broader and more diverse expert finding experiment. As shown

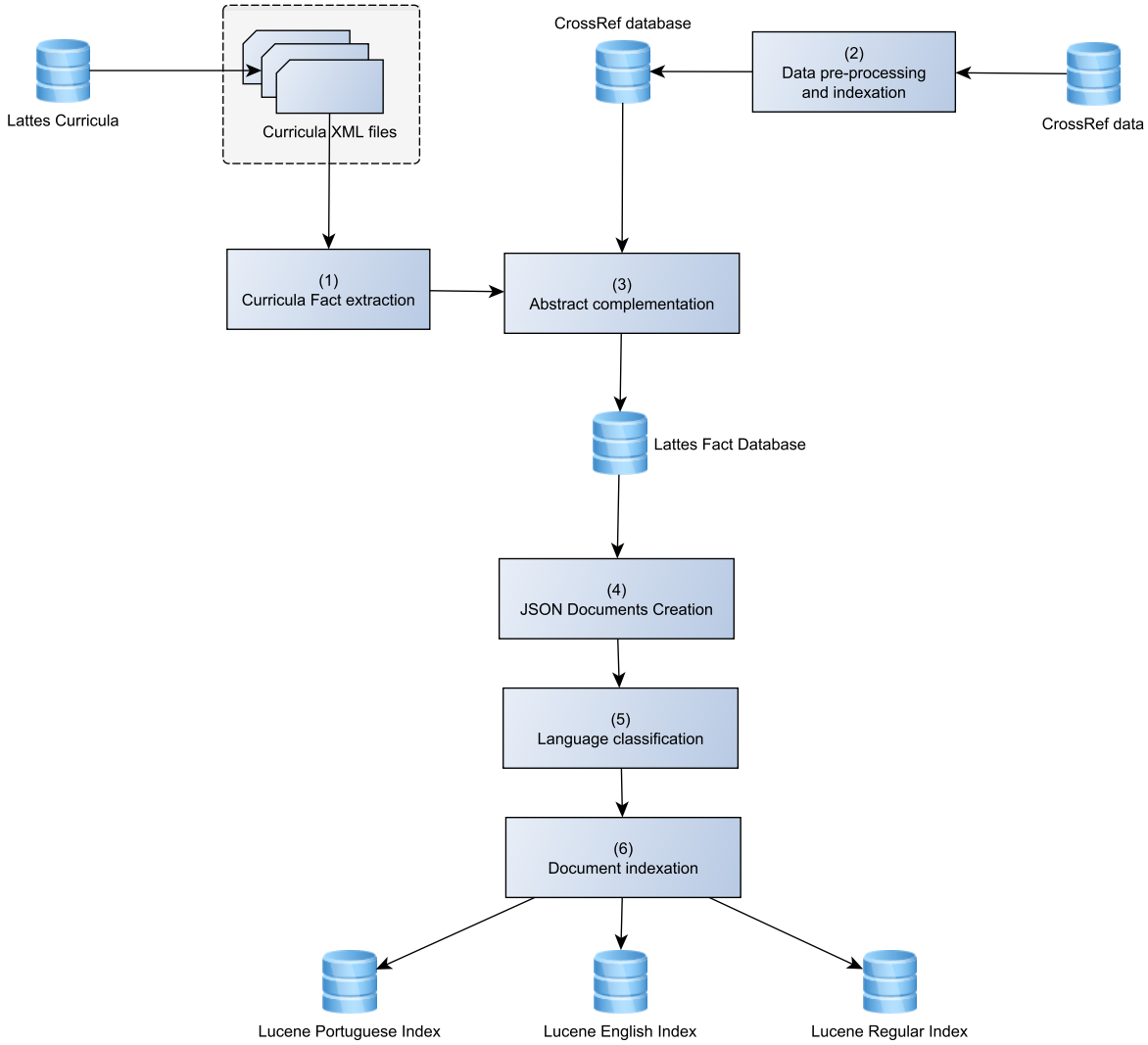


Figure 13 – Dataset preparation

in Chapter 5, the users of our prototype were able to identify candidate experts in other areas besides their own (Computer Science). The curricula were extracted in XML format (**step 1**), which prevented us from dealing with data extraction from an unstructured source, which was not our focus in this work.

The curricula included published articles (both in Journals and Events). Although the platform allows including the Abstract as part of the associated data, most articles in the dataset did not include an Abstract. Thus, to improve the dataset, we developed a tool (**step 2**) to locate the abstracts in CrossRef[2]. It worked as follows: the CrossRef database (a large set of JSON files) was parsed, and the information we required (DOI, Authors, Title, Abstract, Year) was stored in a PostGRES database, indexed by the title, using an n-gram index.

With the Curricula in XML format and the CrossRef database available, we developed another tool (**step 3**) to import the XML files into a PostGRES database,

[2] https://www.crossref.org/

which we will call the *Experion* database. This database was structured with a table for each kind of Fact extracted from the XML curricula: Award, Education, Event Article, Periodic Article, Examining Board, Orientation, and Published Book. The database also contains a table with the list of researchers from which the curricula were extracted. This table contains the researcher's name and a unique ID. Each record on the various Fact-related tables is associated with a record in this researcher's table. Figure 14 introduces an overview of the database schema.

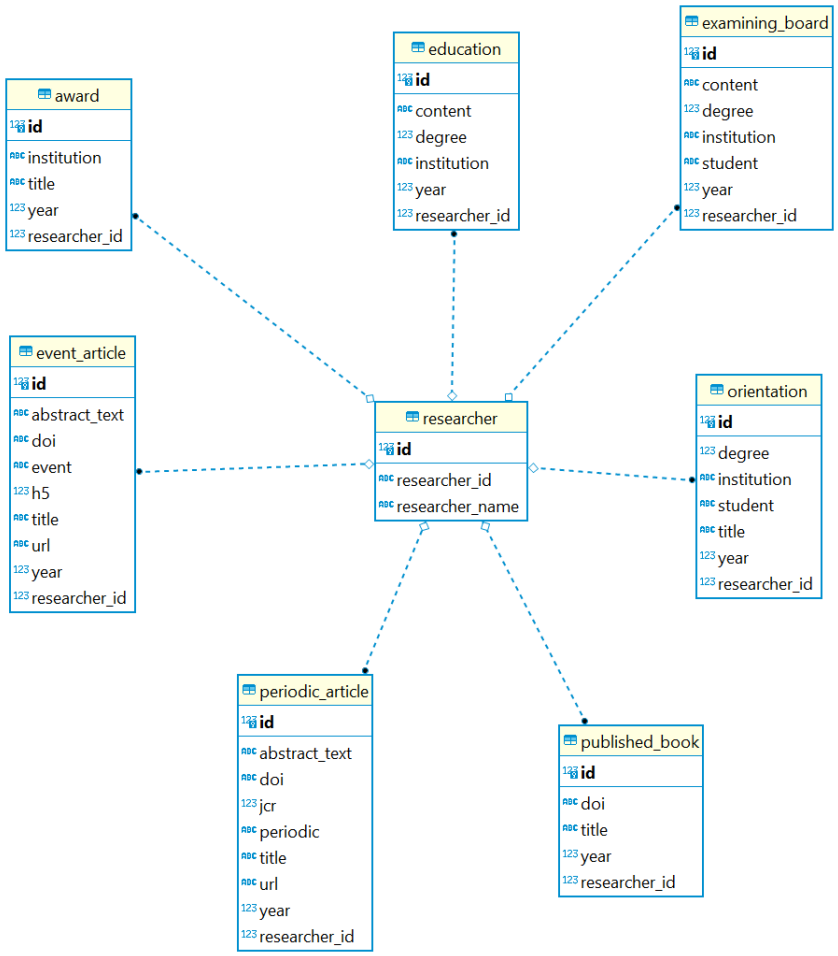


Figure 14 – Experion database schema

After extracting the data and persisting it in the database, the tool tried to locate the missing abstracts for the Periodic and Event Articles. For this, it tried to locate the Article in the Crossref database, using a similarity search based on the title. We adopted a similarity search due to issues with how the titles were written (typos, missing words, and abbreviations, for example). Once the results from the CrossRef database were obtained for a given Article, the tool located in the results which one was associated with the same authors (also using string similarity) and the same year. When it found a record matching these conditions, it updated the Experion database with the abstract.

Once the Experion database was completed, the next step was to generate an inverted index to allow searching for given expertise of interest over the Experion

database. Although such a search could be done over the relational database, an inverted index allowed better performance as well as support for lemmatized search. We elected the Apache Lucene[3] tool to create this inverted index. Lucene allows indexing documents using a text (which can be a set of keywords) and supports lemmatizing in various languages, including Portuguese and English.

To build the inverted index, we developed a tool (**step 4**) that generated a JSON document for each Fact stored in the Experion database. The record contained a unique ID to identify the Fact, the contextual and index dimensions (as established in Section 4), the name, and a unique ID associated with the researcher to whom the evidence was associated. Each of these generated documents was indexed using the text in their associated index dimensions in three Lucene indexes: a regular index (without any kind of lemmatizing), an English lemmatized index, and a Portuguese lemmatized index.

Each document was indexed in the regular and the language indexes. In order to identify the language, we utilized (**step 5**) the LibreTranslate project tool, which, given a text (the contents of the index dimensions from the document), indicated the language. If the language was not English or Portuguese, the document was left only in the regular index (**step 6**). The Lucene indexes were then utilized in the contextualized expertise tool used for the experiments, described in the following.

4.3.2 Implementation

To promote our experiments, we developed an implementation of the Experion framework using the Lattes dataset prepared in the previous section. The system was structured as two modules: a backend and a frontend module. The backend module provides a public REST API consumed by the frontend module.

The **backend** was developed using Django [4], a well-known open-source Python framework. The rest API it provides consists, at the moment, of a single endpoint that allows a parameterized search for given expertise of interest. This endpoint expects the following parameters:

- The keywords which define the expertise of interest. The keywords can be in either English or Portuguese. The backend used the LibreTranslate tool to translate the terms.
- The percentage of top results desired. By result, here we mean the candidate experts. Since we are searching over six thousand curricula using lemmatized keywords, several candidate experts may not be of interest. Thus only the top results are returned. By default, the backend returns the top 10% experts, with the additional rule limiting to 30 experts and at least ten experts. These additional limitations keep the results in a reasonable length.

[3] https://lucene.apache.org/

[4] `https://www.djangoproject.com/`

Once the backend receives a search request, in general terms, it operates as follows:

1. It translates the terms to either English or Portuguese, depending on the input;
2. For each index (regular, Portuguese, and English), it searches for documents matching the input request in the corresponding language;
3. The results from each index are combined (using the unique id associated with each document stored in the indexes) and organized by the expert;
4. The experts are sorted by their result count, and the desired percentage of top candidate experts are selected;
5. The results are structured in a JSON document, where the Facts are grouped by their associated experts and organized by their associated year (the starting year if it spans through more than one year);
6. The resulting JSON document is returned to the client.

In the first and second steps, which translate and search the terms, the backend first uses the LibreTranslate tool to identify the language in which the query was submitted. If it cannot identify the language, it will search only in the regular, non-lemmatized index for the input keywords. If it detects the keywords in English, it will translate them to Portuguese and vice-versa. With the translated terms, it will perform a lemmatized search for the terms in their corresponding language term. For example, if one were to search for "relational databases," the backend would translate it to Portuguese and perform the following searches:

1. **relat databas** in the English index;
2. **banc dado relacional** in the Portuguese index;
3. **relational databases** in the regular, non-lemmatized index.

Some results of the regular index could be the same as the other index. This comes from the fact that all Facts are indexed in the regular index, and those which could have their language identified are also stored in the lemmatized indexes. Thus, the third step of the search process identified and eliminated these duplicates while organizing (grouping) the Facts by their associated researcher and the year.

The **front end** was developed using Angular [5], a well-known open-source Javascript framework. We developed a simple and direct interface (shown in Figure 15), where the user inputs the desired expertise of interest and top percentage of top results (candidate experts) that they want as a result.

After performing the search, an index with the candidate experts, ranked by the number of matching Facts, is presented (Figure 16). This index indicates the number

[5] `https://angular.io/`

of Facts found. By clicking on the name of the experts, the user can view and navigate through the contextualized view of the expertise Facts found associated with that expert (as shown in Figure 17). The index and contextual dimensions from the Facts are displayed in a formatted form (built by the backend in HTML), and the derived dimensions are shown in a highlighted way using *pills*. In the example, the derived dimensions are Level, Location, and Collaboration.

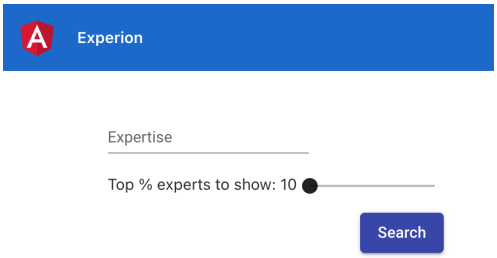


Figure 15 – Experion search input

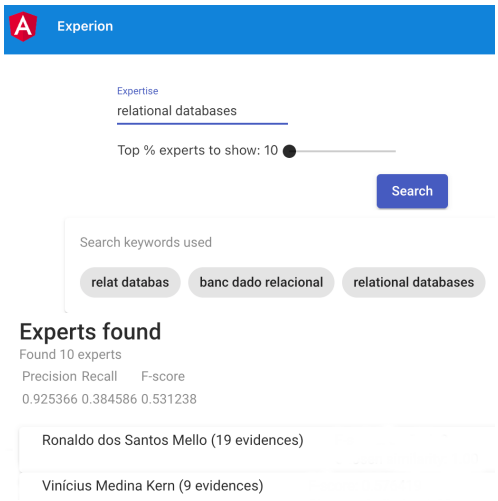


Figure 16 – Experion search index

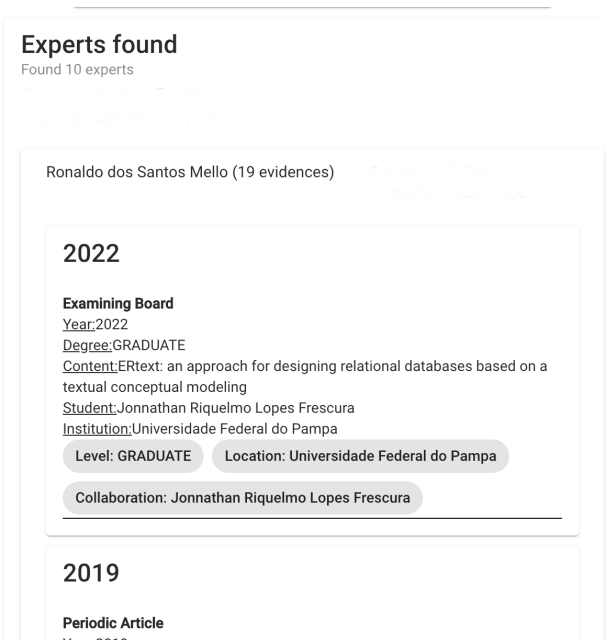


Figure 17 – Experion result navigation

With the tool developed, we could proceed with our experiments, which methodology we describe in the next section.

Question	Evaluation				
	1	**2**	**3**	**4**	**5**
Q1 - How beneficial was the addition of context in the search result?	1	0	4	9	**15**
Q2 - Did the context help to better understand the expertise?	1	2	1	10	**14**
Q3 - The expertise description level was adequate?	0	5	3	**13**	8
Q4 - How understandable were the results without context?	8	**11**	7	3	0
Q5 - How understandable were the results with context?	0	1	6	9	**13**
Q6 - The relevance of the results was understandable without context?	7	6	**12**	4	0
Q7 - The relevance of the results was understandable with context?	0	2	3	**12**	**12**
Q8 - Your satisfaction with the results without context.	8	6	**10**	5	0
Q9 - Your satisfaction with the results with context.	1	3	1	**14**	10

Table 9 – Answers to questions 1 to 9

4.3.3 Methodology

With our developed reference implementation of the Experion framework, to validate its proposal, we chose to promote an unsupervised experiment with a group of users to evaluate the impact of contextualizing an expert finding system. The experiment was composed of the following:

- a brief explanation of the concept of expertise and expertise contextualization,
- a tutorial on how to use the tool, and
- a brief questionnaire to be filled out after using the tool.

Since it was an unsupervised experiment, we did not set a specific time limit to use the tool before filling out the questionnaire. From some initial tests, 20 minutes was a relatively good amount of time, and we suggested such duration in the invitation message. We sent the invitation to researchers, students, and IT personnel of the university's Computer Science and IT departments, using e-mail as the communication medium. After three weeks and two re-invitations, we collected 29 answers to the questionnaire and analyzed the results, as shown in the next section.

4.3.4 Results

The questionnaire presented to the users contained eleven questions labeled Q1 through Q11. The users could answer questions Q1 to Q9 with a value from 1 (worst) to 5 (best), as introduced in Table 9. There were specific options for questions Q10 and Q11, as shown in Table 10.

Analyzing the answers to questions Q1, Q2, and Q3, we see that most people considered adding the context to the results very beneficial, with over 70% of the answers being between values 4 and 5. Questions Q4 and Q5 evaluated whether the user considered adding the context helped them understand how the candidate experts obtained or demonstrated the expertise of interest. As the results demonstrate, over

Question	Evaluation		
	Insufficient	**Sufficient**	**Excessive**
Q10 - The elements used to describe the expertise were: insufficient, sufficient, excessive.	7	21	1

Question	Evaluation	
	Adequate	**Inadequate**
Q11 - A ranking process without considering the context is adequate or inadequate?	10	19

Table 10 – Q10 and Q11 results

65% of the users considered the results without context hard to understand, while over 70% considered the results more understandable when a context was provided.

Recognizing the results' relevance is another crucial goal in expert finding systems - a candidate expert should be relevant to the user's interests. Questions Q6 and Q7 analyze how the users understood the relevance of the results without and with context. Without context, most users (over 85%) considered the results not satisfactory (answers between values 1 and 3), while with context, most users (over 82%) considered the relevance of the results excellent (answers between values 4 and 5).

Questions Q8 and Q9 analyzed the general satisfaction with the results. Here, it is essential to note that our tool was a prototype without focusing on the user interface experience. However, even with such limitations, the general satisfaction was very high when the results introduced context. Over 80% of the users had high satisfaction with the results (answers between 4 and 5).

Since the context description can vary regarding the data presented to the users, we evaluated if the current level of detailing for the expertise context was adequate through questions Q10 and Q11, introduced in Table 10. The tool displays the Facts and context information textually, without additional processing/summarizing. Most users (over 70%) considered the level of detailing sufficient, while only 3% considered them excessive. About 24% considered the detailing insufficient.

Since our tool did not consider the context to rank the candidate experts, we asked the users if they considered that a ranking process without considering the context was adequate. Interestingly, different from what we expected, a reasonable percentage of the users (around 34%) considered the current ranking process satisfactory. That may indicate that the number of Facts per itself is already a good indicator of expertise for raking. Naturally, including the context in the ranking process seems to be a good idea for most users. Some users also provided feedback that some kinds of contexts/Facts should impact the ranking more than others.

As shown in our analysis of the answers from the questionnaire proposed to the users, the majority considered it beneficial to add context to the results, thus validating our hypothesis. Based on our previous research and the results from the experiment,

we could identify that, although Experion was able to generate context as expected, it was not uncommon for this context to lack some expertise evidence (based on some tests we promoted internally during the development of the prototype). Such lack of context can hinder Experion's primary objective, which is to contextualize the evidence well enough for a user to understand. Thus we decided to work on improving context generation, elaborating a method for context injection, introduced Chapter 5.

5 EXPERTISE INJECTION

During our analysis of the results obtained by our application of the Experion Framework in the Lattes Curricula, implemented for the experiment introduced in Chapter 3, we identified an issue with the Lattes Curricula data. Only some expertise evidence had a reasonable amount of context information (contextual dimensions) associated with them due to facts such as the researcher not entering all the required information in his Lattes Curriculum. Such an issue can hinder our proposal of using context to improve the understanding of expert finding systems and may also be an issue with other data sources besides the Lattes Platform.

It became clear that, in order for our proposal to perform accordingly, there is a need to improve the context information present in the data. For such, we developed the concept of *context injection* in the Experion Framework, as an extension to its already defined process, applied before the *Derivators*. Section 5.1 describes our proposed context injection method. An optimization of its process, regarding how its parameters are calculated, is introduced in Section 5.2. An experiment to validate the correctness of the proposed context injection is presented in Section 5.3, and which results are analyzed in 5.3.2. Lastly, Section 5.4 presents a qualitative experiment (an interview) performed with three experts to evaluate our proposal, including context injection.

5.1 CONTEXT INJECTION

To improve the quality of context information associated with expertise evidence, we developed the concept of Context Injection in the Experion Framework. Suppose that an evidence E_a lacks context information. We investigate other related evidence (they could share certain contextual information with evidence E_a, for example) and share the context information with evidence E_a for those we deem applicable.

Our proposed method is completely automated and, as demonstrated by the experiments in Section 5.3, is capable of injecting context with reasonable accuracy. Our method can infer new context for the evidence based on the available context information. It is important to note that the new context is obtained from other evidence and not generated using an external data source. The overall process is shown in Figure 18 and comprises:

1. The expertise evidence found in the expertise retrieval search is separated via their associated candidate expert;
2. For each set of evidence, a graph is generated, including the contextual information from the evidence;
3. A graph-similarity metric is applied, identifying the nodes (expertise evidence) deemed similar by our metrics;

4. Those pieces of evidence deemed similar have their contextual information shared between them.

5.1.1 Graph Generation

The first step of the process is already executed by the Experion Framework, which is to locate the expertise evidence based on search criteria and organize (separate) this evidence (Facts) by expert. A graph is built for each set of evidence, i.e., per candidate expert. The graph generation is completely automatized by the framework. Figure 18 introduces an example of the generated graphs. It contains the following elements as vertices:

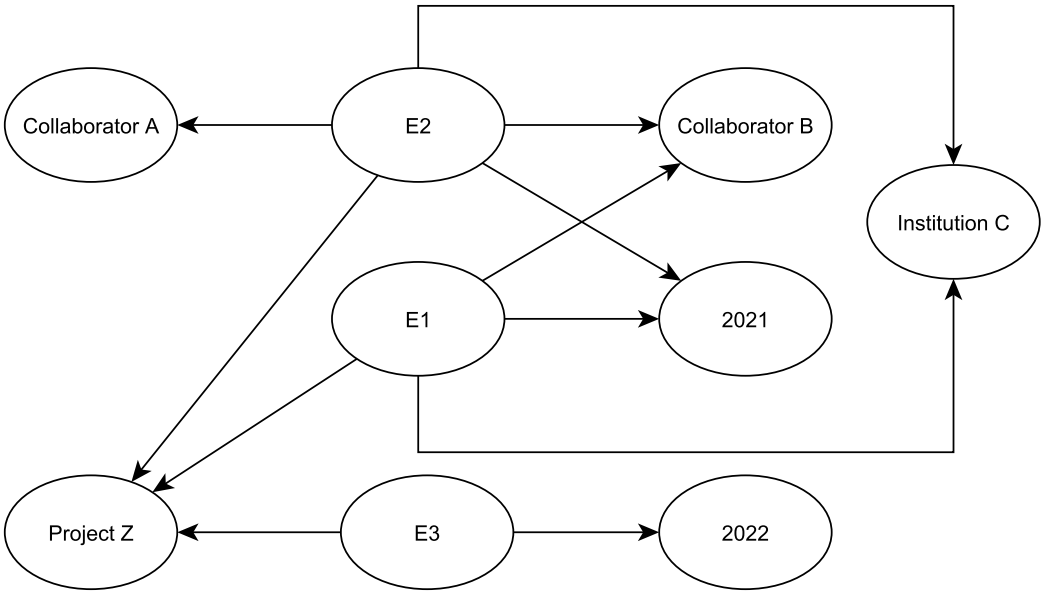


Figure 18 – Context Injection Example Graph

- Year - represents a year where expertise evidence is present. If the evidence is present for several years, all the years are introduced as vertices. We include a single node per year. In the example, we have two years: 2021 and 2022;
- Context information - a node that represents context information. The context information, in this case, is a Derived Dimension value. A single node is created for each Derived Dimension value associated with expertise evidence. If the same value of a Derived Dimension appears more than once in the expertise evidence, only a single vertex is included in the graph. In the example, we have two kinds of context: cooperation (Collaborator A and Collaborator B) and location (Project Z and Institution C);

- Expertise evidence - a node that represents expertise evidence. Every piece of evidence produced by Experion is included in the graph. In Figure 18, we have three pieces of evidence: E1, E2, and E3.

We establish the following unweighted edges in the graph:

- between every Expertise evidence and their associated years and;
- between every Context Information and the Expertise Evidence associated with it.

5.1.2 Graph similarity

To analyze the similarity between the evidence (which are represented as vertices in the graph built in the previous step), we use the SimRank algorithm (JEH; WIDOM, 2002). In general terms, SimRank considers the similarity between two vertices as how many common vertices are related to the two vertices, including vertices indirectly associated (i.e., associated through intermediary vertices). SimRank's result is a table relating the similarity between all the nodes in the graph.

We iterate over this table and locate the highest and lowest similarities found by SimRank. With these limits, we establish a similarity range and find the highest quartile of such range. All evidence pair that has a similarity above this threshold are considered similar and candidate evidence for context injection.

Table 11 – Evidence similarity

	E1	**E2**	**E3**
E1	1.0	0.7204	0.6291
E2	X	1.0	0.6126
E3	X	X	1.0

As an example, let us consider the Graph introduced in Figure 18. Table 11 contains the similarities calculated between the Expertise nodes after applying SimRank for our example. As we can see, the highest similarities for each evidence are: $E_1 -> E_2 = 0.7204$, $E_2 -> E_1 = 0.7204$, $E_3 -> E_1 = 0.6291$. We have a similarity range from [0.6291,0.7204]. The upper quartile, in this case, will be $(0.7204 - (0.7204 - 0.6291)/4) = (0.7204 - 0.02282) = 0.6975$. Thus the only candidate expertise evidence pair for injection would be $E_1 -> E_2$. The other candidate pairs are ignored.

This approach was used based on our experiments, where the similarities varied significantly for each expert/expertise of interest due to different context information availability.

5.1.3 Contextual information sharing

Once we have the similarity range calculated, we iterate through the pairs of similarity between expertise evidence. If a pair is above the threshold found, we inject

the context from the evidence into each other. In our example, evidence E_1 and E_2 will share their context information due to the injection process. Figure 19 demonstrates the resulting graph. The edge with a thicker line indicates the injected context.

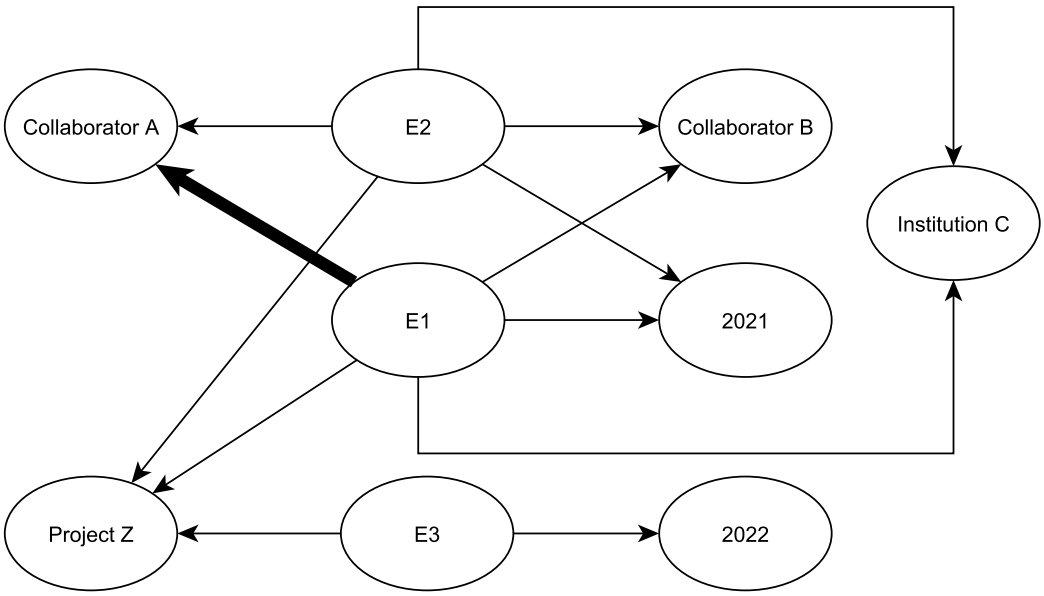


Figure 19 – Context Injection Example Graph - Injected

With this last step, the context injection process is concluded, and the resulting data can be used to generate the Expertise Retrieval results through the Experion Framework.

5.2 OPTIMIZING THE SIMILARITY RANGE

An issue we found during our tests is that the similarity range considered for electing the expertise evidence pairs candidates for context injection impacts the result's quality. Our basic method of using the upper quartile did not produce optimal results in several cases during our experiments. That is directly associated with the difference in available context information between experts and the expertise search executed. Thus, we developed a method to automatically choose the best similarity range to optimize the results.

Our method uses the F-score (RIJSBERGEN, 1979) value and a base ground truth to find the best similarity range. This ground truth is composed of a set of expertise evidence that we know which context information is supposed to be present at the end of the context injection process. Thus, we can calculate the F-score based on the number of missing and wrongly injected context information in the evidence.

Since we cannot have this ground truth pre-calculated for every possible expert and expertise query, we developed a method to create it dynamically. The method works as follows:

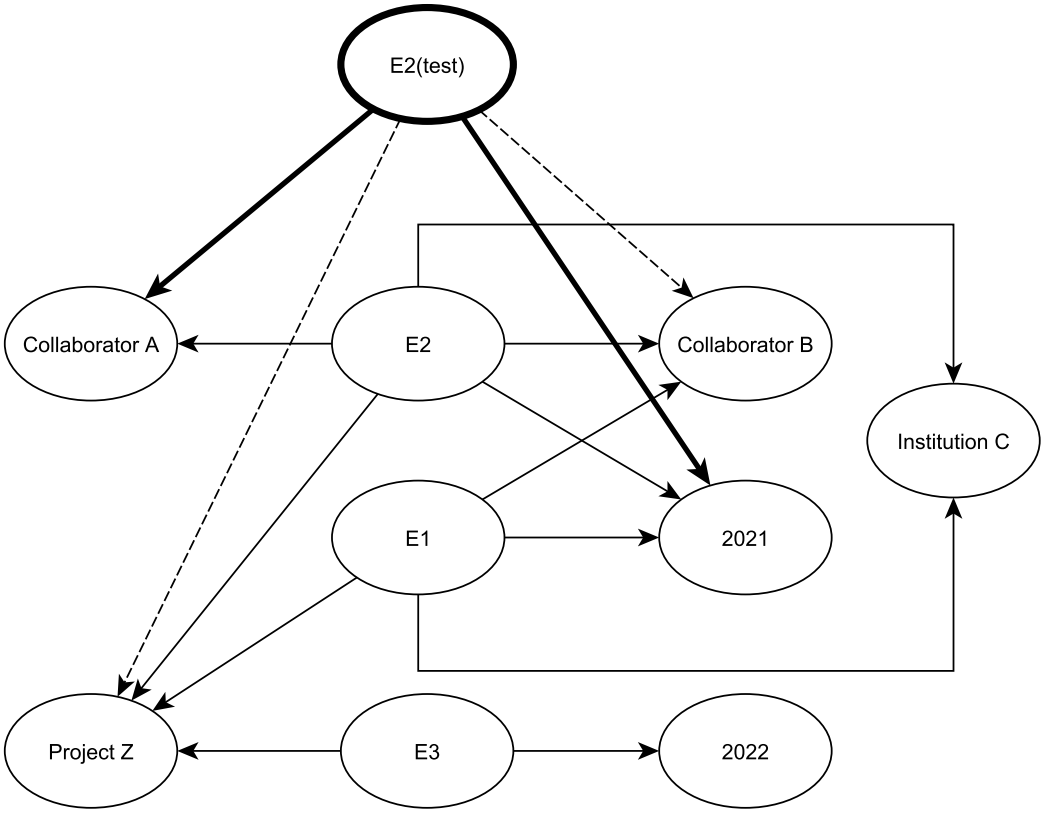


Figure 20 – Context Injection Example Graph - Test evidence

1. First, we separate the expertise evidence per year;
2. We randomly choose half of the expertise evidence for each year and duplicate it. We name these duplicates as *test evidence*.
3. We randomly remove half of their associated context information for each *test evidence*.

Figure 20 demonstrates these steps in our example graph. Initially, we have three pieces of evidence - E1, E2, and E3. Taking half of them means taking a single piece of evidence. Let us suppose that the chosen evidence (since it is a random process) is E2. A new evidence node is created in the graph, named - E2(test), and it is associated with the same contextual information and year as the original evidence E2. After this step, half of these contextual information associations are removed. In our example, the randomly chosen associations were those of Collaborator B and Project Z, which are dashed in Figure 20. The remaining contextual information associated with the test evidence is indicated as thick lines in the figure.

Now, besides the original evidence set, we have additional generated expertise evidence, which we named *test evidence*. Since we generated such test evidence, we know precisely their correct context information. Since we have a ground truth, we can apply the proposed context injection process and then calculate the F-score of the result (considering only the context information initially present in the test evidence).

Using such ground truth, we run several iterations of the context injection process, varying the similarity range in each step. That means that, instead of choosing a range based on the upper quartile of the similarities found by the SimRank method, we try different ranges, for example, between 0.3 to 0.9. For example, suppose that SimRank provides a similarity range for the best matches between pieces of evidence of [0.345, 0.9745]. The iteration process described here injects context between evidence pairs with similarity above the thresholds shown in Table 12.

Table 12 – Evidence similarity ranges

Range	**Minimum similarity**
0.1	0.91155
0.2	0.8486
...	...
0.8	0.4709
0.9	0.40795

For each iteration, the F-Score is calculated for the test evidence. The similarity threshold, which provides the best F-Score, is used to execute the actual context injection.

5.3 EXPERIMENTS

To evaluate our context injection method we developed a reference implementation, improving on the tool shown in Chapter 3. We promoted changes in the backend and front-end modules to allow context injection.

5.3.1 Implementation

In the backend module, we introduced an additional parameter in the query method provided by the REST API that allows specifying a desired F-Score. Per our definition, the desired F-Score will always be 1.0 (the best possible), and if not present, the closest will be chosen. Thus, this parameter assumes a value of 1.0 per default. If specified, the backend will choose the lowest possible F-Score in the injection process above the given value or the closest to it if none above or equal is found.

During the backend process, a new step was introduced to provide the context injection after querying and separating the found expertise evidence by expert. To build the graph and apply the SimRank method, we adopted the NetworkX[1] Python library. This library application is straightforward and returns a table with similarities between the nodes, given the graph. For the resulting JSON from the REST API, additional information was included. The changes include:

[1] `https://networkx.org/`

- For each expert: the f-score found (based on test evidence data), the chosen similarity range (as described previously), and the list of expertise evidence found for the expert. The list of evidence includes the *test evidence*, which is marked accordingly.
- For each piece of evidence: if it suffered context injection, it is marked accordingly, with a confidence score (the similarity calculated between the expertise evidence deemed similar for context injection) also included in the result. The context information present in evidence as the result of the injection process is also marked.

In the front-end module, in the search form, we introduced a field to specify the desired F-Score value, with a default value of 1.0, as explained previously (Figure 21). In the result index (Figure 22) the calculated F-Score for each expert is also shown. For each piece of evidence, if the contextual dimension is injected, it is shown with a different color (blue) - Figure 23 introduces an example. If the user activates the "Show test evidence" (Figure 22), the results include the generated test evidence with a different background. In Figure 24, an example of test evidence is shown.

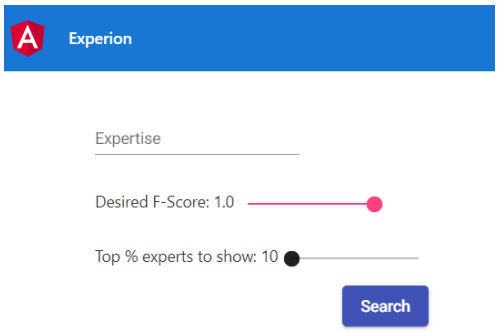


Figure 21 – Experion search interface with F-Score definition

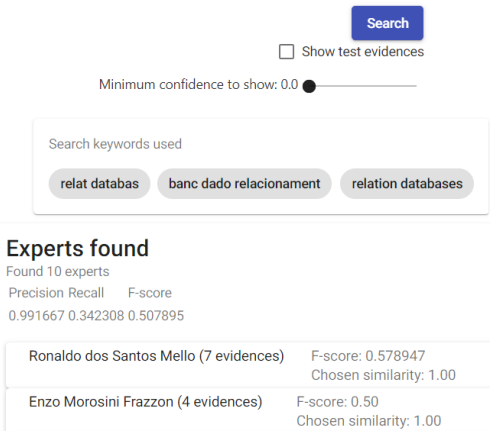


Figure 22 – Experion result index with F-Score values

5.3.2 Context injection performance

Using our reference implementation, we analyzed the performance of our context injection method. Since a human analysis of the results would be time-consuming and very dependent on understanding the results and the context of the expertise evidence by the person analyzing, we opted for an automated approach. That was possible and relatively straight, given that we had already analyzed the quality of the context injection as part of the process to define the similarity range to use.

We executed some expertise searches using the front-end developed and analyzed these queries' precision, recall, and F-score. The resulting data is presented

Figure 23 – Experion result with injected context

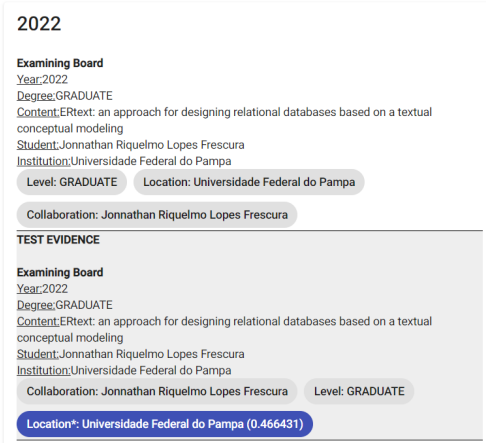


Figure 24 – Experion result with test evidence

Table 13 – Experiment results

Keywords	Precision	Recall	F-Score
database	0.978317	0.474079	0.607093
webforms	0.942486	0.431502	0.563492
artificial intelligence	0.921501	0.591125	0.68876
data mining	0.897541	0.493965	0.618024
crawler	1.000000	0.504762	0.6375
networks	0.854206	0.491019	0.696842

in Table 13. The expertise terms used for the search were random topics in computer science: database, web forms, artificial intelligence, data mining, crawler, and networks.

The table shows that our context injection method presents a good precision value (at least above 0.85), a reasonable recall (around 0.5), and an average F-Score of around 0.6. Such results demonstrate that our method can inject context information with reasonable confidence and is completely automated without fine-tuning depending on the data set.

5.4 EXPERT INTERVIEW

After developing the context injection, we performed another experiment to collect user feedback. Instead of another quantitative, unsupervised experiment, we did a qualitative one. This experiment was composed of an interview with three experts from our institution (UFSC) in the computer science area. The experts were chosen based on their long experience in their fields. Each expert had a different knowledge of expertise retrieval and was interviewed separately. We adopted a semi-structured form to perform the interviews, which took around one hour each and were organized as the following protocol:

1. An initial explanation that included: (i) the concepts of expertise retrieval

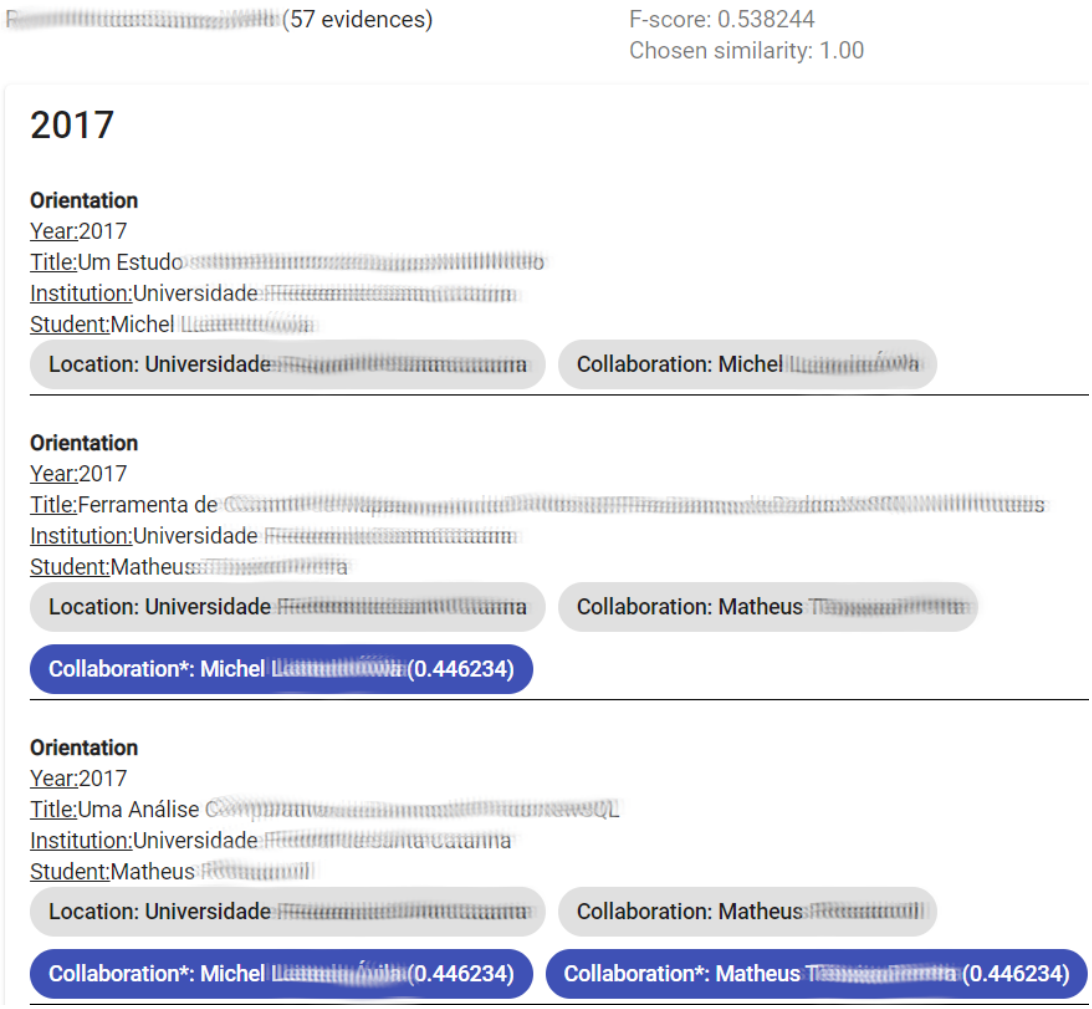


Figure 25 – Experion search details

and expert finding; (ii) a brief overview of the state of current work; (iii) our proposal (contextualization of results) and; (iv) the experiment tool interface and functionality;

2. The expert used the tool freely, searching for expertise of interest - further explanations were provided as needed;
3. During the tool usage, the context injection process was explained, and the expert was asked to identify and analyze context injections.

This protocol focused on evaluating the user experience (UX) with the developed tool. A supervised laboratory test evaluation context was adopted through a prototype, where the evaluation data were obtained through an interview using the developed tool. Since it is a prototype, the focus was on the user perception regarding the injection of context in expert finding results, not the tool's usability. Nonetheless, feedback on the usability was obtained and is included here.

We aimed at a qualitative assessment of the proposed solution, seeking to understand the user's (expert) perception of the system and the benefit of its use. For that

purpose, we observed the use of the system and asked questions to assess the user's perception of the tool(KAPLAN; MAXWELL, 2005). Considering the categorization proposed by Vermeeren et al. (VERMEEREN et al., 2010), our protocol is categorized as follows (only applicable items were considered):

- **Information source:** Specific selection of users
- **Location:** Online on the Web
- **Product development phase:** Functional prototype
- **Period of experience:** Typical test session
- **Type of collected data:** Qualitative

5.4.1 Experts overview

The **first expert**, which we will refer to as *Expert A*, had a basic understanding of expertise retrieval. He promoted several searches with the following keywords: (i) computacao evolutiva; (ii) robotica movel; (iii) algoritmos geneticos; (iv) robotica evolutiva; (v) evolutionary robotics; (vi) multiple aspects trajectory; and (vii) logic synthesis.

The **second expert**, referred to as *Expert B*, had an intermediary knowledge of expertise retrieval. He promoted searches with the following keywords: (i) databases; (ii) nosql; (iii) information retrieval.

The **third expert**, *Expert C*, had no prior knowledge of expertise retrieval. Given the background of this expert in interface design, he focused mainly on the interface elements in the beginning, understanding its structure and how it relates to our proposal concepts. That allowed some very interesting suggestions, which we describe in the following. *Expert C* performed the following searches: (i) teste de software; (ii) blockchain.

5.4.2 Experiment analysis and feedback

The terms used by **Expert A** were mainly very specific, resulting in small sets of evidence. The results produced by the tool included the context from the Lattes Platform, as expected. However, additional context generated through our context injection method was hindered due to the limited available data. Since we use a graph-based approach and the results sets were very small, more data was needed to allow the SimRank algorithm to identify similarities between the Facts. Such an issue happens due to our proposed approach, which uses only data available in the Facts present in the result set and prevents possible false context injection by using the F-Score method. Since little data is available, trying to inject context results in a low F-Score in the test evidence, and thus the algorithm prevents the injections. Nonetheless, since the tool allowed seeing the test evidence and analyzing how the context was injected into them, it became clear to Expert A how the process works.

While analyzing the results, *Expert A* found results from candidate experts from other areas - it was due to the terms used for the search being a hit for the evidence from these candidates. However, according to the expert, some of them were false matches. He could evaluate it due to the contextualization of the expertise that Experion promotes.

Expert A also questioned whether we performed the opposite process: given a researcher, find their expertise. It was explained that this was known as *expert profiling* and was not in the scope of this thesis. Nonetheless, Experion can also be used for profiling and is future work in sight. The expert also suggested using the curricula from researchers in other institutions and allowing a match of experts between the institutions. Similar to expert profiling, the concept of *collaboration suggestion* - a research topic in expertise retrieval - was also explained to *Expert A*.

Expert B used the tool in an alternative way to *Expert A*. Instead of several searches and browsing the results without dwelling on them, *Expert B* analyzed them with extreme interest and cared about the results. Given that the first term of the search resulted in several results (due to its generality) and he was the first-ranked candidate expert, he thoroughly navigated the result. He remembered several activities during this process from his career while analyzing and agreeing to most of the injected context - given the more significant amount of data, the injection process performed better compared to *Expert A*.

Expert C used the tool in a similar way to Expert B. He promoted fewer searchers and analyzed the results with care, interested in seeing which people have been working with the expertise of interest. He noted especially the presence of expertise information from years ago, for which our contextualization, including temporality, provided enough information for him to filter such *candidate experts*.

5.4.3 Suggestions

Expert A suggested including the candidate experts' area of knowledge (information available in the Lattes Platform) in the results. That would allow a user to quickly filter those of interest by looking at the areas associated with the candidate experts. He also questioned if the framework could use other data sources besides the Lattes Platform. Similar to expert profiling, we detailed future work which involves using other data sources present in the university (undergraduate and graduate management systems) as data sources to improve the information available.

Regarding the context injection process, due to the random aspects of it, *Expert A* also suggested an improvement to the process: perform several injections for the same data set, in which each run would randomly choose different evidence and context to process and, after these runs, use the average f-score to perform the final injection and also use the variation in f-score between the runs to calculate a variability degree.

This suggestion was very interesting and will be incorporated into future work.

During the analysis of the context injection, **Expert B** suggested allowing navigation of the results using the associated context. This idea came from an injection of a collaborator that could not be recognized. It was later found to be due to an examination board where this person participated. Allowing the user to use the injected context as a hypergraph, for example, to find all evidence associated with the same context, would improve the user experience in the expert opinion, which we agree and include as future work.

Another interesting fact found during the interview of *Expert B* was that, when searching for a more general term (information retrieval), one expert he expected did not appear. Discussing the fact, we understood that it was because the expert published several topics in the area of information retrieval. However, these specific keywords probably do not appear in the evidence. While analyzing this issue (after the interview), a possible solution would be to use the support of some classification system, such as the ACM Computing Classification System [2], to improve the keywords used for the expertise evidence retrieval automatically. That will be included in future work as well.

The first suggestion by **Expert C** was to allow filtering of the results by year. That was due to his first search (teste de software) presenting some candidate experts who worked with the topic several years ago. Thus they are not relevant anymore. Such an issue reinforces the importance of contextualization and temporalization. That is similar to what *Expert A* experienced while using the tool.

As in the case of *Expert A*, *Expert C* also suggested that having the area of knowledge associated with the candidate expert would improve the analysis of the results. However, instead of associating each expert with the area, Expert C suggested having a filter on the interface where the undesired areas of knowledge could be discarded from the results. Thus he could look only into the results of areas that he considers compatible with his interest while searching for an expert.

Following the idea of result filtering, allowing a filter by the context associated with the results was also suggested by *Expert C*. The proposition was that, given a list of all contexts associated with the results, the user could filter out those in which he is not interested. For example, removing events/journals users know are not interesting or related to the expertise of interest being searched.

Another suggestion by *Expert C* was considering the evidence type (article, book examination board) as a context. This suggestion came from the fact that, in Expert C's opinion, someone who just participated in an examination board about a subject is not an expert for the purpose he was looking at the results (to find which people are working on his topics of interest). While analyzing this suggestion, we concluded that allowing the evidence type as a context would be interesting. Since we will implement a filter

[2] `https://dl.acm.org/ccs`

by context information, considering the evidence type as context would allow filtering them. For example, someone looking for a person to compose an examination board would be more interested in people that worked on previous boards on the subject of interest, while another looking for a researcher would be more interested in previous participation in projects.

5.4.4 Closing remarks

At the end of the interview, **Expert A** mentioned how important it is to improve communication and collaboration between researchers, especially in the institution's context. Thus the proposed framework could be a valuable tool in this process.

Like *Expert A*, **Expert B** mentioned how important it is to improve communication and collaboration between researchers. He was receptive to the possibility of using the Lattes Curricula from all institutions in Brazil and locating experts in other institutions. He also suggested help in the developed tool, explaining the interface usage and concepts.

Similar to the prior Experts, **Expert C** mentioned how important it is to improve the communication and collaboration between researchers. For that, having a tool to find people working on given expertise topics is very important. He was receptive to the possibility of using the Lattes Curricula from all institutions in Brazil and locating experts in other institutions.

5.5 CONCLUSION

In this chapter, we introduced a context injection method based solely on existing data to improve the contextualization of results in expert finding. We interviewed three experts to analyze its impact and collect a qualitative analysis of the Experion framework. To allow a comprehensive and diverse analysis of our proposal, even with reduced sample size, we elected experts with very different backgrounds and interests. *Expert A* was very interested in the structure and functionality of our framework and its application over the Lattes platform. Due to his prior knowledge of the subject, *Expert B* became more interested in how the contextualization and context injection were processed and analyzed the results more critically than the other Experts. *Expert C* had a different interest while using the tool - he made some suggestions considering how the interface could be optimized to improve the analysis and understanding of the proposed contextualization.

Naturally, the study case used in this experiment, which was applying the framework over the Lattes Platform, introduces some simplifications to the process that should be considered in future work. Due to its semi-structured format, the Lattes curriculum facilitates extracting context data from its Facts. It may be a complex issue with other data sources that do not have such associated meta-data - for example, web

pages. Also, the context generation provided in this experiment was a basic description of the DerivedDimensions and did not consider using more advanced techniques, such as natural language description. Such a possibility is the focus of future work as well. Nevertheless, these simplifications do not limit or hinder the framework's validity. Its structure as a *black-box* framework with a *contract-based* approach allows further expansion and exchange of components between different implementations.

Although the three Experts had different approaches and interests, it was common sense that our proposal of introducing Facts and context to an expert finding system promoted a better understanding of its results. Such feedback, together with our experiments in Chapter 4, validated this work's hypothesis. In Chapter 6, based on the feedback from the interviewed experts, we introduce some exciting possibilities for future work and improvements.

6 CONCLUSION

We developed two novel contributions to Expertise Retrieval in this work: a faceted taxonomy and the Experion framework. The taxonomy proposed classifies Expertise Retrieval work over several perspectives, such as what kind of data source is used, which techniques are used, and what is the final application (expert finding, expert profiling, among others). The Experion framework allowed contextualizing expertise evidence for expert finding.

Initially, based on the taxonomy we introduced, an extensive survey on the state of the art of expertise retrieval was promoted. Based on this survey, several open issues in Expertise Retrieval were identified and analyzed. Among the open issues, we focused our two related issues: contextualization and explanation of the results, and, through contextualization, tackled the issue of explanation of results.

To provide contextualization in expert finding, we developed a Framework called Experion. This framework comprises a set of entities (concepts) - Entity, Fact, Dimension, and Context - and functions to build the expertise context - Derivator Functions and Context Builders. We introduced an application of Experion in the context of expert finding in the Academia, describing in detail the implementation of the framework in such context. Our implementation is publicly available as well.

Expert finding is a crucial application for expertise retrieval. In a context where expertise evidence is being generated in large volumes daily, it becomes a vital tool for finding experts for a given task. Such a task may seem trivial at first (find the expert who works on a given topic) but has two key challenges: extracting the expertise evidence and understanding it. The first challenge has been well addressed in the literature, with several methods to locate and extract expertise information from existing data. On the other hand, the second challenge is still an open issue that we focused on addressing in this work.

Using our proposed framework Experion, any expert finding system can extract the context information associated with a set of expertise evidence. Such functionality provides the user of the expert finding systems with a better understanding of the expert finding system results. Nonetheless, only some expertise evidence contains adequate context information. Considering this issue, we developed a context injection method introduced in this work.

Since, to the best of our knowledge, we are proposing a novel approach to improve the results of expert finding systems, based on our hypothesis that a context improves the understanding of such results, there was a need to validate the hypothesis. As such evaluation is a subjective issue, we adopted a user-based evaluation of an expert finding system with context integrated into the results. As shown in our analysis of the answers from the questionnaire proposed to the users, the majority considered it

beneficial to add context to the results, thus validating our hypothesis.

To improve the contextualization of the results, we proposed a context injection method to increase the context available in the results. Our proposal injects context on a given expertise evidence based on the evidence to which it is related. Our approach is entirely automatic, does not require external information, and is capable of self-tuning by using test evidence. Our experiments demonstrate that it performs reasonably well, with an average F-score around 0.6.

Both our framework proposal and the context injection method were subject to a qualitative experiment, where three experts were interviewed and request to use our tool and analyze the results. These interviews provided several ideas for future work, among other we had planned. We have the following future work on contextualizing expert finding systems results:

- Include the general area of knowledge of the expertise evidence as a ContextualDimension.
- Expand the data source used by including curricula from other institutions and other platforms as well, besides Lattes, and analyze the framework performance.
- Extend the context injection process by executing several rounds and choosing the best result - since we randomly choose the test evidence, the randomicity may significantly impact the results. We will also consider how to improve and analyze the efficacy of our proposed context injection method.
- Testing alternative methods to calculate the similarity between expertise evidence (besides SimRank) and define which contexts should be injected. We will use these alternative methods to promote qualitative tests with several experts, analyzing their feedback on the different context injection methods.
- Elaborating a ranking proposal for the results considering the context information.
- Providing alternative navigation in the results, through a hypergraph over the contextual information, allowing to find all evidence related to a given ContextualDimension, for example.
- Improving the search query made by the user by using related terms from the same domain.
- Allowing filtering of the results by the Contextual information.
- Establishing standard implementations for the framework components, allowing easy usage of other data sources besides the Lattes data source and creating common-use shared libraries.

- Developing new ways to describe the contexts found by the framework using a story-based approach, using a more natural and human-like form.
- Using other data sources to validate our framework generality, such as Google Scholar, LinkedIn, and ResearchGate is a future research topic.

We also intend to apply our framework to other applications in Expertise Retrieval, such as Expert Profiling and Collaboration Recommendation.

This work contribution resulted in four publications - all of them in Qualis-qualified [1] venues. The two principal publications were: (i) a survey about expertise retrieval, which introduced our faceted taxonomy, published in the ACM Computing Surveys (Qualis A1) (GONÇALVES; DORNELES, C. F., 2019); (ii) and a paper published in the WebMedia'22 event (A4) (GONÇALVES; DORNELES, C. F., 2022), which presented the Experion framework. We also published a short paper in the SBBD'22 event (A4) (GONÇALVES; DORNELES, C., 2022) about identifying named entities in Lattes Curricula and a paper in the WTDBD at the SBBD'21 event (GONÇALVES; DORNELES, C., 2021).

Four undergraduate co-orientations were performed as well on the following topics: (i) grouping Lattes Curricula by Knowledge Area Affinity with Temporality (SILVA, 2016); (ii) grouping researchers by co-authoring (COLONETTI, 2016); (iii) temporal identification of expertise (PIZZINATTO, 2019) and; (iv) collecting and comparing publication data from online sources (GoogleScholar, DBLP, ResearchGate) with Lattes Curricula (BRANCO, 2018).

[1] According to https://ppgcc.github.io/discentesPPGCC/pt-BR/qualis/

REFERENCES

ABITEBOUL, Serge; BUNEMAN, Peter; SUCIU, Dan. **Data on the Web: From Relations to Semistructured Data and XML**. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 2000. ISBN 1-55860-622-X.

ALARFAJ, F.; KRUSCHWITZ, U.; FOX, C. An adaptive window-size approach for expert-finding. English. In: CEUR Workshop Proceedings. [S.l.: s.n.], 2013. v. 986, p. 76–79.

ANALYTICS, Clarivate. **Journal Citation Reports**. [S.l.: s.n.], 2017. [Online; accessed 2-February-2017]. Available from: `http://about.jcr.incites.thomsonreuters.com/`.

ARAKI, Masataka; KATSURAI, Marie; OHMUKAI, Ikki; TAKEDA, Hideaki. Interdisciplinary collaborator recommendation based on research content similarity. **IEICE Transactions on Information and Systems**, v. 100, n. 4, p. 1–8, 2017.

B, Liangming Pan; WANG, Zhigang; LI, Juanzi; TANG, Jie. Domain Specific Cross-Lingual Knowledge Linking Based on Similarity Flooding. v. 7091, p. 426–438, 2011.

BALOG, Krisztian; AZZOPARDI, Leif; RIJKE, Maarten de. A language modeling framework for expert finding. **Information Processing & Management**, v. 45, n. 1, p. 1–19, 2009. ISSN 0306-4573.

BALOG, Krisztian; AZZOPARDI, Leif; RIJKE, Maarten de. 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. Seattle, Washington, USA: ACM, 2006. (SIGIR '06), p. 43–50.

BALOG, Krisztian; DE RIJKE, Maarten. Determining Expert Profiles (with an Application to Expert Finding). In: PROCEEDINGS of the 20th International Joint Conference on Artifical Intelligence. Hyderabad, India: Morgan Kaufmann Publishers Inc., 2007. (IJCAI'07), p. 2657–2662.

BALOG, Krisztian; FANG, Yi; RIJKE, Maarten de; SERDYUKOV, Pavel; SI, Luo. Expertise Retrieval. **Foundations and Trends® in Information Retrieval**, v. 6, 2–3, p. 127–256, 2012. ISSN 1554-0669.

BALOG, Krisztian; RIJKE, Maarten de. Finding Experts and Their Eetails in e-Mail Corpora. In: PROCEEDINGS of the 15th International Conference on World Wide Web. Edinburgh, Scotland: ACM, 2006. (WWW '06), p. 1035–1036.

BALOG, Krisztian; RIJKE, Maarten de. Finding Similar Experts. In: PROCEEDINGS of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. Amsterdam, The Netherlands: ACM, 2007. (SIGIR '07), p. 821–822.

BAN, Z.; LIU, L. CICPV: A New Academic Expert Search Model. In: 2016 IEEE 30th International Conference on Advanced Information Networking and Applications (AINA). [S.l.: s.n.], Mar. 2016. P. 47–52.

BERENDSEN, Richard; RIJKE, Maarten de; BALOG, Krisztian; BOGERS, Toine; BOSCH, Antal van den. On the assessment of expertise profiles. **Journal of the American Society for Information Science and Technology**, v. 64, n. 10, p. 2024–2044, 2013. ISSN 1532-2890.

BERGMAN, Michael K. White Paper: The Deep Web. Surfacing Hidden Value. **The Journal of Electronic Publishing**, v. 7, n. 1, online, Aug. 2001.

BHANU, M.; CHANDRA, J. Exploiting response patterns for identifying topical experts in StackOverflow. In: 2016 Eleventh International Conference on Digital Information Management (ICDIM). [S.l.: s.n.], Sept. 2016. P. 139–144.

BIZER, Christian; HEATH, Tom; IDEHEN, Kingsley; BERNERS-LEE, Tim. Linked Data on the Web (LDOW2008). In: PROCEEDINGS of the 17th International Conference on World Wide Web. Beijing, China: ACM, 2008. (WWW '08), p. 1265–1266.

BLEI, David M. Probabilistic Topic Models. **Commun. ACM**, ACM, New York, NY, USA, v. 55, n. 4, p. 77–84, Apr. 2012. ISSN 0001-0782.

BLEI, David M. Probabilistic Topic Models. **Commun. ACM**, ACM, New York, NY, USA, v. 55, n. 4, p. 77–84, Apr. 2012. ISSN 0001-0782.

BLEI, David M.; NG, Andrew Y.; JORDAN, Michael I. Latent Dirichlet Allocation. **J. Mach. Learn. Res.**, JMLR.org, v. 3, p. 993–1022, Mar. 2003. ISSN 1532-4435.

BOEVA, Veselka; BONEVA, Liliana; TSIPORKOVA, Elena. Semantic-Aware Expert Partitioning. In: **Artificial Intelligence: Methodology, Systems, and Applications: 16th International Conference, AIMSA 2014, Varna, Bulgaria, September 11-13, 2014. Proceedings**. Ed. by Gennady Agre, Pascal Hitzler, Adila A. Krisnadhi and Sergei O. Kuznetsov. Cham: Springer International Publishing, 2014. P. 13–24. ISBN 978-3-319-10554-3.

BOLELLI, Levent; ERTEKIN, Şeyda; GILES, C. Lee. Topic and Trend Detection in Text Collections Using Latent Dirichlet Allocation. In: PROCEEDINGS of the 31th European Conference on IR Research on Advances in Information Retrieval. Toulouse, France: Springer-Verlag, 2009. (ECIR '09). P. 776–780. ISBN 978-3-642-00957-0.

BRANCO, Arthur Machado. **Ferramenta para coleta e comparação de dados de publicações acadêmicas dos professores com o currículo lattes**. [S.l.], 2018. Available from: `https://repositorio.ufsc.br/handle/123456789/192306`.

BUDALAKOTI, S.; DEANGELIS, D.; BARBER, K. S. Expertise Modeling and Recommendation in Online Question and Answer Forums. v. 4, p. 481–488, Aug. 2009.

CABANAC, Guillaume. Accuracy of Inter-researcher Similarity Measures Based on Topical and Social Clues. **Scientometrics**, Springer-Verlag New York, Inc., Secaucus, NJ, USA, v. 87, n. 3, p. 597–620, June 2011. ISSN 0138-9130.

CAMPBELL, Christopher S.; MAGLIO, Paul P.; COZZI, Alex; DOM, Byron. Expertise Identification Using Email Communications. In: PROCEEDINGS of the Twelfth International Conference on Information and Knowledge Management. New Orleans, LA, USA: ACM, 2003. (CIKM '03), p. 528–531.

CHA, Youngchul; CHANG, Keng-hao; BOMMAGANTI, Hari; CHEN, Ye; YAN, Tak; BI, Bin; CHO, Junghoo. A Universal Topic Framework (UniZ) and Its Application in Online Search. In: PROCEEDINGS of the 30th Annual ACM Symposium on Applied Computing. Salamanca, Spain: ACM, 2015. (SAC '15), p. 1078–1085.

CHAIWANAROM, Paweena; LURSINSAP, Chidchanok. Collaborator recommendation in interdisciplinary computer science using degrees of collaborative forces, temporal evolution of research interest, and comparative seniority status. **Knowledge-Based Systems**, v. 75, p. 161–172, 2015. ISSN 0950-7051.

CHEN, Hung-Hsuan; GOU, Liang; ZHANG, Xiaolong; GILES, Clyde Lee. CollabSeer: A Search Engine for Collaboration Discovery. In: PROCEEDINGS of the 11th Annual International ACM/IEEE Joint Conference on Digital Libraries. Ottawa, Ontario, Canada: ACM, 2011. (JCDL '11), p. 231–240.

CHEN, Hung-Hsuan; TREERATPITUK, Pucktada; MITRA, Prasenjit; GILES, C. Lee. CSSeer: An Expert Recommendation System Based on CiteseerX. In: PROCEEDINGS of the 13th ACM/IEEE-CS Joint Conference on Digital Libraries. Indianapolis, Indiana, USA: ACM, 2013. (JCDL '13), p. 381–382.

CHEN, Xu; ZHOU, Mingyuan; CARIN, Lawrence. The Contextual Focused Topic Model. In: PROCEEDINGS of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Beijing, China: ACM, 2012. (KDD '12), p. 96–104.

CHENG, X.; ZHU, S.; CHEN, G.; SU, S. Exploiting User Feedback for Expert Finding in Community Question Answering, p. 295–302, Nov. 2015.

COCARASCU, Oana; MCLEAN, Andrew; FRENCH, Paul; TONI, Francesca. **An Explanatory Query-Based Framework for Exploring Academic Expertise**. [S.l.]: arXiv, 2021.

COHEN, Sara; EBEL, Lior. Recommending Collaborators Using Keywords. In: PROCEEDINGS of the 22Nd International Conference on World Wide Web. Rio de Janeiro, Brazil: ACM, 2013. (WWW '13 Companion), p. 959–962.

COLONETTI, Gabriela Bussolo. **Agrupando pesquisadores por coautoria de publicações segundo currículo Lattes**. [S.l.], 2016. Available from: `https://repositorio.ufsc.br/handle/123456789/171398`.

CUMMINS, Ronan; LALMAS, Mounia; O'RIORDAN, Colm. Learning Aggregation Functions for Expert Search. In: PROCEEDINGS of the 2010 Conference on ECAI 2010: 19th European Conference on Artificial Intelligence. Amsterdam, The Netherlands, The Netherlands: IOS Press, 2010. P. 535–540.

DARGAHI NOBARI, Arash; SOTUDEH GHAREBAGH, Sajad; NESHATI, Mahmood. Skill Translation Models in Expert Finding. **Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval - SIGIR '17**, p. 1057–1060, 2017.

DAUD, Ali. Using time topic modeling for semantics-based dynamic research interest finding. **Knowledge-Based Systems**, v. 26, p. 154–163, 2012. ISSN 0950-7051.

DAUD, Ali; LI, Juanzi; ZHOU, Lizhu; MUHAMMAD, Faqir. Exploiting Temporal Authors Interests via Temporal-Author-Topic Modeling. In: PROCEEDINGS of the 5th International Conference on Advanced Data Mining and Applications. Beijing, China: Springer-Verlag, 2009. (ADMA '09), p. 435–443.

DAVOODI, Elnaz; KIANMEHR, Keivan; AFSHARCHI, Mohsen. A semantic social network-based expert recommender system. **Applied Intelligence**, v. 39, n. 1, p. 1–13, 2013. ISSN 1573-7497.

DEHGHAN, Mahdi; BIABANI, Maryam; ABIN, Ahmad Ali. Temporal expert profiling: With an application to T-shaped expert finding. **Information Processing & Management**, Elsevier BV, v. 56, n. 3, p. 1067–1079, May 2019.

DENG, Hongbo; HAN, Jiawei; LYU, Michael R.; KING, Irwin. Modeling and Exploiting Heterogeneous Bibliographic Networks for Expertise Ranking. In: PROCEEDINGS of the 12th ACM/IEEE-CS Joint Conference on Digital Libraries. Washington, DC, USA: ACM, 2012. (JCDL '12), p. 71–80.

DIJK, David van; TSAGKIAS, Manos; RIJKE, Maarten de. 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**, p. 995–998, 2015.

DOM, Byron; EIRON, Iris; COZZI, Alex; ZHANG, Yi. Graph-based Ranking Algorithms for e-Mail Expertise Analysis. ACM, San Diego, California, p. 42–48, 2003.

DONG, X. L.; SRIVASTAVA, D. Big data integration. In: 2013 IEEE 29th International Conference on Data Engineering (ICDE). [S.l.: s.n.], Apr. 2013. P. 1245–1248.

DU, Jianguang; JIANG, Jing; SONG, Dandan; LIAO, Lejian. Topic Modeling with Document Relative Similarities. In: PROCEEDINGS of the 24th International Conference on Artificial Intelligence. Buenos Aires, Argentina: AAAI Press, 2015. (IJCAI'15), p. 3469–3475.

DUONG, Trong Hai; NGUYEN, Ngoc Thanh; JO, Geun Sik. Constructing and Mining a Semantic-based Academic Social Network. **J. Intell. Fuzzy Syst.**, IOS Press,

Amsterdam, The Netherlands, The Netherlands, v. 21, n. 3, p. 197–207, Aug. 2010. ISSN 1064-1246.

FANG, Yi; GODAVARTHY, Archana. Modeling the Dynamics of Personal Expertise. In: PROCEEDINGS of the 37th International ACM SIGIR Conference on Research & Development in Information Retrieval. Gold Coast, Queensland, Australia: ACM, 2014. (SIGIR '14), p. 1107–1110.

FANG, Yi; SI, Luo; MATHUR, Aditya P. 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. Geneva, Switzerland: ACM, 2010. (SIGIR '10), p. 683–690.

FAYAD, Mohamed; SCHMIDT, Douglas Clark. Object-Oriented Application Frameworks. **Commun. ACM**, Association for Computing Machinery, New York, NY, USA, v. 40, n. 10, p. 32–38, Oct. 1997. ISSN 0001-0782.

FAZEL-ZARANDI, Maryam; FOX, Mark S. Constructing Expert Profiles over Time for Skills Management and Expert Finding. In: PROCEEDINGS of the 11th International Conference on Knowledge Management and Knowledge Technologies. Graz, Austria: ACM, 2011. (i-KNOW '11), 5:1–5:6.

GAO, Shengxiang; LI, Xian; YU, Zhengtao; QIN, Yu; ZHANG, Yang. Combining paper cooperative network and topic model for expert topic analysis and extraction. **Neurocomputing**, v. 257, p. 136–143, 2017. Machine Learning and Signal Processing for Big Multimedia Analysis. ISSN 0925-2312.

GOLLAPALLI, Sujatha Das; MITRA, Prasenjit; GILES, C. Lee. Ranking Experts Using Author-document-topic Graphs. In: PROCEEDINGS of the 13th ACM/IEEE-CS Joint Conference on Digital Libraries. Indianapolis, Indiana, USA: ACM, 2013. (JCDL '13), p. 87–96.

GOLLAPALLI, Sujatha Das; MITRA, Prasenjit; GILES, C. Lee. Similar Researcher Search in Academic Environments. In: PROCEEDINGS of the 12th ACM/IEEE-CS Joint Conference on Digital Libraries. Washington, DC, USA: ACM, 2012. (JCDL '12), p. 167–170.

GONÇALVES, Rodrigo; DORNELES, Carina. Experion: A framework for contextualizing evidence in expert finding. In: ANAIS Estendidos do XXXVI Simpósio Brasileiro de Bancos de Dados. Rio de Janeiro: SBC, 2021. P. 113–119.

GONÇALVES, Rodrigo; DORNELES, Carina. Identifying named entity from researcher curricula. In: ANAIS do XXXVII Simpósio Brasileiro de Bancos de Dados. Búzios: SBC, 2022. P. 427–432.

GONÇALVES, Rodrigo; DORNELES, Carina Friedrich. Automated Expertise Retrieval: A Taxonomy-Based Survey and Open Issues. **ACM Comput. Surv.**, Association for Computing Machinery, New York, NY, USA, v. 52, n. 5, Sept. 2019. ISSN 0360-0300.

GONÇALVES, Rodrigo; DORNELES, Carina Friedrich. Context Injection in Expert Finding. In: PROCEEDINGS of the Brazilian Symposium on Multimedia and the Web. Curitiba, Brazil: Association for Computing Machinery, 2022. (WebMedia '22), p. 168–177.

GRIFFITHS, Thomas L.; STEYVERS, Mark. Finding scientific topics. **Proceedings of the National Academy of Sciences**, v. 101, suppl 1, p. 5228–5235, 2004. eprint: http://www.pnas.org/content/101/suppl_1/5228.full.pdf.

HASHEMI, Seyyed Hadi; NESHATI, Mahmood; BEIGY, Hamid. Expertise Retrieval in Bibliographic Network: A Topic Dominance Learning Approach. In: PROCEEDINGS of the 22Nd ACM International Conference on Information & Knowledge Management. San Francisco, California, USA: ACM, 2013. (CIKM '13), p. 1117–1126.

HE, Qi; CHEN, Bi; PEI, Jian; QIU, Baojun; MITRA, Prasenjit; GILES, Lee. Detecting Topic Evolution in Scientific Literature: How Can Citations Help? In: PROCEEDINGS of the 18th ACM Conference on Information and Knowledge Management. Hong Kong, China: ACM, 2009. (CIKM '09), p. 957–966.

HIEMSTRA, Djoerd. **Using language models for information retrieval**. [S.l.: s.n.], 2001. ISBN 9075296053.

HIRSCH, J. E. An index to quantify an individual's scientific research output. **Proceedings of the National Academy of Sciences of the United States of America**, v. 102, n. 46, p. 16569–16572, 2005. eprint: http://www.pnas.org/content/102/46/16569.full.pdf.

HOANG, Nguyen Le; KHOA, Pham Vu Dang; PHUC, Do. Predicting preferred topics of authors based on co-authorship network. In: THE 2013 RIVF International Conference on Computing Communication Technologies - Research, Innovation, and Vision for Future (RIVF). [S.l.: s.n.], Nov. 2013. P. 70–75.

HOFMANN, Katja; BALOG, Krisztian; BOGERS, Toine; RIJKE, Maarten de. Contextual factors for finding similar experts. **Journal of the American Society for Information Science and Technology**, Wiley Subscription Services, Inc., A Wiley Company, v. 61, n. 5, p. 994–1014, 2010. ISSN 1532-2890.

HUANG, S.; TANG, Y.; TANG, F.; LI, J. Link prediction based on time-varied weight in co-authorship network. In: PROCEEDINGS of the 2014 IEEE 18th International Conference on Computer Supported Cooperative Work in Design (CSCWD). [S.l.: s.n.], May 2014. P. 706–709.

J, Ganesh; GANGULY, Soumyajit; GUPTA, Manish; VARMA, Vasudeva; PUDI, Vikram. Author2Vec: Learning Author Representations by Combining Content and Link Information. International World Wide Web Conferences Steering Committee, Montréal, Québec, Canada, p. 49–50, 2016.

JAMEEL, Shoaib; LAM, Wai. An N-Gram Topic Model for Time-Stamped Documents. In: **Advances in Information Retrieval: 35th European Conference on IR Research, ECIR 2013, Moscow, Russia, March 24-27, 2013. Proceedings**. Ed. by Pavel Serdyukov, Pavel Braslavski, Sergei O. Kuznetsov, Jaap Kamps, Stefan Rüger, Eugene Agichtein, Ilya Segalovich and Emine Yilmaz. Berlin, Heidelberg: Springer Berlin Heidelberg, 2013. P. 292–304. ISBN 978-3-642-36973-5.

JAMEEL, Shoaib; LAM, Wai. An Unsupervised Topic Segmentation Model Incorporating Word Order. In: PROCEEDINGS of the 36th International ACM SIGIR Conference on Research and Development in Information Retrieval. Dublin, Ireland: ACM, 2013. (SIGIR '13), p. 203–212.

JEH, Glen; WIDOM, Jennifer. SimRank: A Measure of Structural-Context Similarity. In: PROCEEDINGS of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Edmonton, Alberta, Canada: Association for Computing Machinery, 2002. (KDD '02), p. 538–543.

JIANG, Y.; LI, X.; MENG, W. DiscWord: Learning Discriminative Topics. In: 2014 IEEE/WIC/ACM International Joint Conferences on Web Intelligence (WI) and Intelligent Agent Technologies (IAT). [S.l.: s.n.], Aug. 2014. v. 2, p. 63–70.

JIN, Jian; GENG, Qian; ZHAO, Qian; ZHANG, Lixue. Integrating the Trend of Research Interest for Reviewer Assignment. In: PROCEEDINGS of the 26th International Conference on World Wide Web Companion. Perth, Australia: International World Wide Web Conferences Steering Committee, 2017. (WWW '17 Companion), p. 1233–1241.

JO, Yookyung; HOPCROFT, John E.; LAGOZE, Carl. The Web of Topics: Discovering the Topology of Topic Evolution in a Corpus. In: PROCEEDINGS of the 20th International Conference on World Wide Web. Hyderabad, India: ACM, 2011. (WWW '11), p. 257–266.

JOHRI, Nikhil; ROTH, Dan; TU, Yuancheng. Experts' Retrieval with Multiword-enhanced Author Topic Model. In: PROCEEDINGS of the NAACL HLT 2010 Workshop on Semantic Search. Los Angeles, California: Association for Computational Linguistics, 2010. (SS '10), p. 10–18.

KAMSIANG, Nawarat; SENIVONGSE, Twittie. An Ontology-Based Methodology for Building and Matching Researchers' Profiles. In: **IAENG Transactions on Engineering Technologies: Special Issue of the World Congress on Engineering and Computer Science 2012**. Ed. by Haeng Kon Kim, Sio-Iong Ao, Mahyar A. Amouzegar and Burghard B. Rieger. Dordrecht: Springer Netherlands, 2014. P. 455–468. ISBN 978-94-007-6818-5.

KAPLAN, Bonnie; MAXWELL, Joseph A. Qualitative Research Methods for Evaluating Computer Information Systems. In: **Evaluating the Organizational Impact of Healthcare Information Systems**. Ed. by James G. Anderson and Carolyn E. Aydin. New York, NY: Springer New York, 2005. P. 30–55. ISBN 978-0-387-30329-1.

KARIMZADEHGAN, Maryam; WHITE, Ryen W.; RICHARDSON, Matthew. Enhancing Expert Finding Using Organizational Hierarchies. In: PROCEEDINGS of the 31th European Conference on IR Research on Advances in Information Retrieval. Toulouse, France: Springer-Verlag, 2009. (ECIR '09), p. 177–188.

KATARIA, Saurabh; MITRA, Prasenjit; CARAGEA, Cornelia; GILES, C. Lee. Context Sensitive Topic Models for Author Influence in Document Networks. In:

PROCEEDINGS of the Twenty-Second International Joint Conference on Artificial Intelligence - Volume Volume Three. Barcelona, Catalonia, Spain: AAAI Press, 2011. (IJCAI'11), p. 2274–2280.

KAWAMAE, Noriaki. Latent Interest-topic Model: Finding the Causal Relationships Behind Dyadic Data. ACM, Toronto, ON, Canada, p. 649–658, 2010.

KAWAMAE, Noriaki. 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. Maui, Hawaii, USA: ACM, 2012. (CIKM '12), p. 2065–2069.

KAYA, Mehmet; ALHAJJ, Reda. Development of multidimensional academic information networks with a novel data cube based modeling method. **Information Sciences**, Elsevier Inc., v. 265, p. 211–224, 2014. ISSN 0020-0255.

KOH, Yun Sing; DOBBIE, Gillian. Indirect Weighted Association Rules Mining for Academic Network Collaboration Recommendations. In: PROCEEDINGS of the Tenth Australasian Data Mining Conference - Volume 134. Sydney, Australia: Australian Computer Society, Inc., 2012. (AusDM '12), p. 167–173.

KONG, Xiangjie; JIANG, Huizhen; BEKELE, Teshome Megersa; WANG, Wei; XU, Zhenzhen. 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. Perth, Australia: International World Wide Web Conferences Steering Committee, 2017. (WWW '17 Companion), p. 1371–1377.

KONG, Xiangjie; JIANG, Huizhen; YANG, Zhuo; XU, Zhenzhen; XIA, Feng; TOLBA, Amr. Exploiting Publication Contents and Collaboration Networks for Collaborator Recommendation. **PLOS ONE**, Public Library of Science, v. 11, n. 2, p. 1–13, Feb. 2016.

KOU, Yue; SHEN, Derong; XU, Hongbin; LIN, Menger; YU, Ge; NIE, Tiezheng. Two-level interactive identification and derivation of topic clusters in complex networks. **World Wide Web**, v. 18, n. 4, p. 1093–1122, 2015. ISSN 1573-1413.

KUMAR, Akshi; JAIN, Abha. An Algorithmic Framework for Collaborative Interest Group Construction. In: **Recent Trends in Networks and Communications:**

International Conferences, NeCoM 2010, WiMoN 2010, WeST 2010, Chennai, India, July 23-25, 2010. Proceedings. Ed. by Natarajan Meghanathan, Selma Boumerdassi, Nabendu Chaki and Dhinaharan Nagamalai. Berlin, Heidelberg: Springer Berlin Heidelberg, 2010. P. 500–508. ISBN 978-3-642-14493-6.

KUMAR, Varun; PEDANEKAR, Niranjan. Mining Shapes of Expertise in Online Social Q&A Communities. In: 7. PROCEEDINGS of the 19th ACM Conference on Computer Supported Cooperative Work and Social Computing Companion. San Francisco, California, USA: ACM, 2016. (CSCW '16 Companion, 7), p. 317–320.

KUNDU, Dipankar; MANDAL, Deba Prasad. Formulation of a hybrid expertise retrieval system in community question answering services. **Applied Intelligence**, Springer Science and Business Media LLC, v. 49, n. 2, p. 463–477, Sept. 2018.

LA ROBERTIE, B. de; PITARCH, Y.; TAKASU, A.; TESTE, O. Identifying Authoritative Researchers in Digital Libraries Using External a Priori Knowledge. In: PROCEEDINGS of the Symposium on Applied Computing. Marrakech, Morocco: ACM, 2017. (SAC '17), p. 1017–1022.

LAGOZE, Carl; VAN DE SOMPEL, Herbert. The Open Archives Initiative: Building a Low-barrier Interoperability Framework. In: PROCEEDINGS of the 1st ACM/IEEE-CS Joint Conference on Digital Libraries. Roanoke, Virginia, USA: ACM, 2001. (JCDL '01), p. 54–62.

LATIF, A.; AFZAL, M. T.; TOCHTERMANN, K. Constructing experts profiles from Linked Open Data. In: 2010 6th International Conference on Emerging Technologies (ICET). [S.l.: s.n.], Oct. 2010. P. 33–38.

LEY, Michael. DBLP: Some Lessons Learned. **Proc. VLDB Endow.**, VLDB Endowment, v. 2, n. 2, p. 1493–1500, Aug. 2009. ISSN 2150-8097.

LI, Chunshan; CHEUNG, William K.; YE, Yunming; ZHANG, Xiaofeng; CHU, Dianhui; LI, Xin. The Author-Topic-Community model for author interest profiling and community discovery. **Knowledge and Information Systems**, v. 44, n. 2, p. 359–383, 2015. ISSN 0219-3116.

LI, Jing; XIA, Feng; WANG, Wei; CHEN, Zhen; ASABERE, Nana Yaw; JIANG, Huizhen. ACRec: A Co-authorship Based Random Walk Model for Academic Collaboration

Recommendation. In: PROCEEDINGS of the 23rd International Conference on World Wide Web. Seoul, Korea: ACM, 2014. (WWW '14 Companion), p. 1209–1214.

LI, Wen; EICKHOFF, Carsten; VRIES, Arjen P. de. Probabilistic Local Expert Retrieval. In: LECTURE Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics). [S.l.: s.n.], 2016. v. 9626. P. 227–239. ISBN 9783319306704. arXiv: 1601.02376.

LI, Y.; MA, S.; ZHANG, Y.; HUANG, R. Expertise Network Discovery via Topic and Link Analysis in Online Communities. In: 2012 IEEE 12th International Conference on Advanced Learning Technologies. [S.l.: s.n.], July 2012. P. 311–315.

LI, Y.; TANG, J. Expertise Search in a Time-Varying Social Network. In: 2008 The Ninth International Conference on Web-Age Information Management. [S.l.: s.n.], July 2008. P. 293–300.

LIANG, Shangsong. Unsupervised Semantic Generative Adversarial Networks for Expert Retrieval. In: THE World Wide Web Conference on - WWW '19. [S.l.]: ACM Press, 2019.

LIANG, Shangsong; RIJKE, Maarten de. Formal language models for finding groups of experts. **Information Processing and Management**, Elsevier Ltd, v. 52, n. 4, p. 529–549, 2016. ISSN 03064573.

LIMA, Rennan C.; SANTOS, Rodrygo L. T. On Extractive Summarization for Profile-centric Neural Expert Search in Academia. In: PROCEEDINGS of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval. [S.l.]: ACM, July 2022.

LIN, Lili; XU, Zhuoming; DING, Ying; LIU, Xiaozhong. Finding topic-level experts in scholarly networks. **Scientometrics**, v. 97, n. 3, p. 797–819, 2013. ISSN 1588-2861.

LIU, Dong; WANG, Li; ZHENG, Jianhua; NING, Ke; ZHANG, Liang-Jie. Influence Analysis Based Expert Finding Model and Its Applications in Enterprise Social Network. In: 2013 IEEE International Conference on Services Computing. [S.l.: s.n.], June 2013. P. 368–375.

LIU, Jingyuan; LIU, Debing; YAN, Xingyu; DONG, Li; ZENG, Ting; ZHANG, Yutao; TANG, Jie. AMiner-mini: A People Search Engine for University. In: PROCEEDINGS of

the 23rd ACM International Conference on Conference on Information and Knowledge Management. Shanghai, China: ACM, 2014. (CIKM '14), p. 2069–2071.

LIU, P.; LIU, K.; LIU, J. Ontology-Based Expertise Matching System Within Academia, p. 5431–5434, Sept. 2007. ISSN 2161-9646.

LIU, Ping; CURSON, Jayne; DEW, Peter. Exploring RDF for Expertise Matching within an Organizational Memory. In: **Advanced Information Systems Engineering: 14th International Conference, CAiSE 2002 Toronto, Canada, May 27–31, 2002 Proceedings**. Ed. by Anne Banks Pidduck, M. Tamer Ozsu, John Mylopoulos and Carson C. Woo. Berlin, Heidelberg: Springer Berlin Heidelberg, 2002. P. 100–116. ISBN 978-3-540-47961-1.

LIU, Ping; CURSON, Jayne; DEW, Peter. Use of RDF for expertise matching within academia. **Knowledge and Information Systems**, v. 8, n. 1, p. 103–130, 2005. ISSN 02191377.

LIU, Xiaomo; WANG, G. Alan; JOHRI, Aditya; ZHOU, Mi; FAN, Weiguo. Harnessing Global Expertise: A Comparative Study of Expertise Profiling Methods for Online Communities. **Information Systems Frontiers**, Kluwer Academic Publishers, Hingham, MA, USA, v. 16, n. 4, p. 715–727, Sept. 2014. ISSN 1387-3326.

LIU, Zhu; LI, Kan; QU, Dacheng. Knowledge Graph Based Question Routing for Community Question Answering. **Neural Information Processing**, v. 10638, p. 721–730, 2017.

LUONG, Ngoc Tu; NGUYEN, Tuong Tri; JUNG, Jason J; HWANG, Dosam. Discovering Co-author Relationship in Bibliographic Data Using Similarity Measures and Random Walk Model. In: NGUYEN, Ngoc Thanh; TRAWIŃSKI, Bogdan; KOSALA, Raymond (Eds.). **Intelligent Information and Database Systems**. Cham: Springer International Publishing, 2015. P. 127–136.

M WEBER, Griffin. **Professional Networking and Expertise Mining for Research Collaboration**. May 2019. Available from: `http://profiles.catalyst.harvard.edu/`. Visited on: 17 May 2019.

MACDONALD, Craig; OUNIS, Iadh. Searching for expertise: Experiments with the voting model. **The Computer Journal**, v. 52, n. 7, p. 729–748, 2009. ISSN 0010-4620.

MANGARAVITE, Vitor; SANTOS, Rodrygo L.T. 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. Pisa, Italy: ACM, 2016. (SIGIR '16), p. 925–928.

MANGARAVITE, Vitor; SANTOS, Rodrygo L.T.; RIBEIRO, Isac S.; GONÇALVES, Marcos André; LAENDER, Alberto H.F. The LExR Collection for Expertise Retrieval in Academia. In: PROCEEDINGS of the 39th International ACM SIGIR Conference on Research and Development in Information Retrieval. Pisa, Italy: ACM, 2016. (SIGIR '16), p. 721–724.

MANNING, Christopher D.; RAGHAVAN, Prabhakar; SCHÜTZE, Hinrich. **Introduction to Information Retrieval**. New York, NY, USA: Cambridge University Press, 2008. ISBN 0521865719, 9780521865715.

MANNING, Christopher D.; RAGHAVAN, Prabhakar; SCHÜTZE, Hinrich. **Introduction to Information Retrieval**. New York, NY, USA: Cambridge University Press, 2008. ISBN 0521865719.

MEYER, Bertrand. Contract-Driven Development. In: DWYER, Matthew B.; LOPES, Antónia (Eds.). **Fundamental Approaches to Software Engineering**. Berlin, Heidelberg: Springer Berlin Heidelberg, 2007. P. 11–11.

MIMNO, David; MCCALLUM, Andrew. Expertise Modeling for Matching Papers with Reviewers. In: PROCEEDINGS of the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. San Jose, California, USA: ACM, 2007. (KDD '07), p. 500–509.

MOU, Haikun; GENG, Qian; JIN, Jian; CHEN, Chong. 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). [S.l.: s.n.], 2015. v. 9460. P. 83–95. ISBN 9783319289397.

MUMTAZ, Sara; RODRIGUEZ, Carlos; BENATALLAH, Boualem. Expert2Vec: Experts Representation in Community Question Answering for Question Routing. In: ADVANCED Information Systems Engineering. [S.l.]: Springer International Publishing, 2019. P. 213–229.

NAVEED, Nasir; SIZOV, Sergej; STAAB, Steffen. ATTention: Understanding Authors and Topics in Context of Temporal Evolution. In: **Advances in Information Retrieval: 33rd European Conference on IR Research, ECIR 2011, Dublin, Ireland, April 18-21, 2011. Proceedings**. Ed. by Paul Clough, Colum Foley, Cathal Gurrin, Gareth J. F. Jones, Wessel Kraaij, Hyowon Lee and Vanessa Mudoch. Berlin, Heidelberg: Springer Berlin Heidelberg, 2011. P. 733–737. ISBN 978-3-642-20161-5.

NESHATI, Mahmood; BEIGY, Hamid; HIEMSTRA, Djoerd. Expert group formation using facility location analysis. **Information Processing and Management**, Elsevier Ltd, v. 50, n. 2, p. 361–383, 2014. ISSN 03064573.

NESHATI, Mahmood; FALLAHNEJAD, Zohreh; BEIGY, Hamid. On dynamicity of expert finding in community question answering. **Information Processing & Management**, Elsevier Ltd, v. 53, n. 5, p. 1026–1042, 2017. ISSN 0306-4573.

NESHATI, Mahmood; HIEMSTRA, Djoerd; ASGARI, Ehsaneddin; BEIGY, Hamid. Integration of scientific and social networks. **World Wide Web**, v. 17, n. 5, p. 1051–1079, 2014. ISSN 1386145X.

OSBORNE, Francesco; MOTTA, Enrico; MULHOLLAND, Paul. Exploring Scholarly Data with Rexplore. In: **The Semantic Web – ISWC 2013: 12th International Semantic Web Conference, Sydney, NSW, Australia, October 21-25, 2013, Proceedings, Part I**. Ed. by Harith Alani. Berlin, Heidelberg: Springer Berlin Heidelberg, 2013. P. 460–477. ISBN 978-3-642-41335-3.

PAL, Aditya. Discovering Experts Across Multiple Domains. In: PROCEEDINGS of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval. Santiago, Chile: ACM, 2015. (SIGIR '15), p. 923–926.

PARADA, Gustavo A.; CEBALLOS, Hector G.; CANTU, Francisco J.; RODRIGUEZ-ACEVES, Lucia. Recommending Intra-institutional Scientific Collaboration Through Coauthorship Network Visualization. In: PROCEEDINGS of the 2013 Workshop on Computational Scientometrics: Theory & Applications. San Francisco, California, USA: ACM, 2013. (CompSci '13), p. 7–12.

PENG, T.; ZHANG, D.; LIU, X.; WANG, S.; ZUO, W. Central Author Mining from Co-authorship Network. In: 2013 Sixth International Symposium on Computational Intelligence and Design. [S.l.: s.n.], Oct. 2013. v. 1, p. 228–232.

PIZZINATTO, Luiz Eduardo. **Exper.te: um framework para identificação temporal da expertise**. [S.l.], 2019. Available from: https://repositorio.ufsc.br/handle/123456789/202507.

PUNNARUT, Ravikarn; SRIHAREE, Gridaphat. A Researcher Expertise Search System Using Ontology-based Data Mining. In: PROCEEDINGS of the Seventh Asia-Pacific Conference on Conceptual Modelling - Volume 110. Brisbane, Australia: Australian Computer Society, Inc., 2010. (APCCM '10), p. 71–78.

RIBEIRO, Isac S.; SANTOS, Rodrygo L.T.; GONÇALVES, Marcos A.; LAENDER, Alberto H.F. On Tag Recommendation for Expertise Profiling: A Case Study in the Scientific Domain. In: PROCEEDINGS of the Eighth ACM International Conference on Web Search and Data Mining. Shanghai, China: ACM, 2015. (WSDM '15), p. 189–198.

RIJSBERGEN, Cornelis Joost van. **Information Retrieval**. 2nd. [S.l.]: Butterworth-Heinemann, 1979.

RYBAK, Jan; BALOG, Krisztian; NØRVÅG, Kjetil. Temporal Expertise Profiling. In: **Advances in Information Retrieval: 36th European Conference on IR Research, ECIR 2014, Amsterdam, The Netherlands, April 13-16, 2014. Proceedings**. Ed. by Maarten de Rijke, Tom Kenter, Arjen P. de Vries, ChengXiang Zhai, Franciska de Jong, Kira Radinsky and Katja Hofmann. Cham: Springer International Publishing, 2014. P. 540–546. ISBN 978-3-319-06028-6.

SATELI, Bahar; LÖFFLER, Felicitas; KÖNIG-RIES, Birgitta; WITTE, René. ScholarLens: Extracting competences from research publications for the automatic generation of semantic user profiles. **PeerJ Computer Science**, PeerJ Inc., v. 3, e121, 7 2017. ISSN 2376-5992.

SERDYUKOV, Pavel; RODE, Henning; HIEMSTRA, Djoerd. Modeling Multi-step Relevance Propagation for Expert Finding. In: PROCEEDINGS of the 17th ACM Conference on Information and Knowledge Management. Napa Valley, California, USA: ACM, 2008. (CIKM '08), p. 1133–1142.

SERDYUKOV, Pavel; TAYLOR, Mike; VINAY, Vishwa; RICHARDSON, Matthew; WHITE, Ryen W. Automatic People Tagging for Expertise Profiling in the Enterprise. In: PROCEEDINGS of the 33rd European Conference on Advances in Information Retrieval. Dublin, Ireland: Springer-Verlag, 2011. (ECIR'11), p. 399–410.

SHI, Chuan; KONG, Xiangnan; YU, Philip S.; XIE, Sihong; WU, Bin. Relevance Search in Heterogeneous Networks. ACM, Berlin, Germany, p. 180–191, 2012.

SILVA, Jaime Mendes da. **tCALC: Agrupamento de Currículos Lattes por Afinidade de Áreas de Conhecimento Considerando Temporalidade**. [S.l.], 2016. Available from: `https://repositorio.ufsc.br/handle/123456789/171420`.

SINGH, Jasmeet; GUPTA, Vishal. Text Stemming: Approaches, Applications, and Challenges. **ACM Comput. Surv.**, ACM, New York, NY, USA, v. 49, n. 3, 45:1–45:46, Sept. 2016. ISSN 0360-0300.

SMIRNOVA, Elena; BALOG, Krisztian. A User-oriented Model for Expert Finding. In: PROCEEDINGS of the 33rd European Conference on Advances in Information Retrieval. Dublin, Ireland: Springer-Verlag, 2011. (ECIR'11), p. 580–592.

STEYVERS, Mark; SMYTH, Padhraic; ROSEN-ZVI, Michal; GRIFFITHS, Thomas. Probabilistic Author-topic Models for Information Discovery. In: PROCEEDINGS of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Seattle, WA, USA: ACM, 2004. (KDD '04), p. 306–315.

SUN, Yizhou; BARBER, Rick; GUPTA, Manish; AGGARWAL, Charu C.; HAN, Jiawei. Co-author Relationship Prediction in Heterogeneous Bibliographic Networks. In: PROCEEDINGS of the 2011 International Conference on Advances in Social Networks Analysis and Mining. Washington, DC, USA: IEEE Computer Society, 2011. (ASONAM '11), p. 121–128.

TANG, Jie. AMiner: Toward Understanding Big Scholar Data. In: PROCEEDINGS of the Ninth ACM International Conference on Web Search and Data Mining. San Francisco, California, USA: ACM, 2016. (WSDM '16), p. 467–467.

TANG, Jie; WU, Sen; SUN, Jimeng; SU, Hang. Cross-domain Collaboration Recommendation. In: PROCEEDINGS of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Beijing, China: ACM, 2012. (KDD '12), p. 1285–1293.

TANG, Jie; YAO, Limin; ZHANG, Duo; ZHANG, Jing. A Combination Approach to Web User Profiling. **ACM Trans. Knowl. Discov. Data**, ACM, New York, NY, USA, v. 5, n. 1, 2:1–2:44, Dec. 2010. ISSN 1556-4681.

TANG, Jie; ZHANG, Jing; JIN, Ruoming; YANG, Zi; CAI, Keke; ZHANG, Li; SU, Zhong. Topic level expertise search over heterogeneous networks. **Machine Learning**, v. 82, n. 2, p. 211–237, 2011. ISSN 1573-0565.

TANG, Jie; ZHANG, Jing; YAO, Limin; LI, Juanzi; ZHANG, Li; SU, Zhong. ArnetMiner: Extraction and Mining of Academic Social Networks. In: PROCEEDINGS of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Las Vegas, Nevada, USA: ACM, 2008. (KDD '08), p. 990–998.

THO, Quan Thanh; HUI, Siu Cheng; FONG, A. C. M. A Web mining approach for finding expertise in research areas. In: PROCEEDINGS. 2003 International Conference on Cyberworlds. [S.l.: s.n.], Dec. 2003. P. 310–317.

TU, Yuancheng; JOHRI, Nikhil; ROTH, Dan; HOCKENMAIER, Julia. Citation Author Topic Model in Expert Search. Association for Computational Linguistics, Beijing, China, p. 1265–1273, 2010.

VAN GYSEL, Christophe; RIJKE, Maarten de; KANOULAS, Evangelos. Structural Regularities in Text-based Entity Vector Spaces, 2017. arXiv: 1707.07930.

VAN GYSEL, Christophe; RIJKE, Maarten de; WORRING, Marcel. Unsupervised, Efficient and Semantic Expertise Retrieval. International World Wide Web Conferences Steering Committee, Montréal, Québec, Canada, p. 1069–1079, 2016.

VERMEEREN, Arnold P. O. S.; LAW, Effie Lai-Chong; ROTO, Virpi; OBRIST, Marianna; HOONHOUT, Jettie; VÄÄNÄNEN-VAINIO-MATTILA, Kaisa. User Experience Evaluation Methods: Current State and Development Needs. In: PROCEEDINGS of the 6th Nordic Conference on Human-Computer Interaction: Extending Boundaries. Reykjavik, Iceland: Association for Computing Machinery, 2010. (NordiCHI '10), p. 521–530.

VERTOMMEN, Joris; JANSSENS, Frizo; DE MOOR, Bart; DUFLOU, Joost R. Multiple-vector User Profiles in Support of Knowledge Sharing. **Inf. Sci.**, Elsevier Science Inc., New York, NY, USA, v. 178, n. 17, p. 3333–3346, Sept. 2008. ISSN 0020-0255.

WANG, Jianwen; HU, Xiaohua; TU, Xinhui; HE, Tingting. Author-conference Topic-connection Model for Academic Network Search. In: PROCEEDINGS of the 21st

ACM International Conference on Information and Knowledge Management. Maui, Hawaii, USA: ACM, 2012. (CIKM '12), p. 2179–2183.

WANG, Xiaolong; ZHAI, Chengxiang; ROTH, Dan. 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. Chicago, Illinois, USA: ACM, 2013. (KDD '13), p. 1115–1123.

WU, Hao; CHELMIS, Charalampos; SORATHIA, Vikram; ZHANG, Yinuo; PATRI, Om Prasad; PRASANNA, Viktor K. Enriching employee ontology for enterprises with knowledge discovery from social networks. **Proceedings of the 2013 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining - ASONAM '13**, p. 1315–1322, 2013.

XIE, Xiaoqin; LI, Yijia; ZHANG, Zhiqiang; PAN, Haiwei; HAN, Shuai. A Topic-Specific Contextual Expert Finding Method in Social Network. In: **Web Technologies and Applications: 18th Asia-Pacific Web Conference, APWeb 2016, Suzhou, China, September 23-25, 2016. Proceedings, Part I**. Ed. by Feifei Li, Kyuseok Shim, Kai Zheng and Guanfeng Liu. Cham: Springer International Publishing, 2016. P. 292–303. ISBN 978-3-319-45814-4.

XU, Shuo; SHI, Qingwei; QIAO, Xiaodong; ZHU, Lijun; JUNG, Hanmin; LEE, Seungwoo; CHOI, Sung-Pil. Author-Topic over Time (AToT): A Dynamic Users' Interest Model. In: PARK, James J. (Jong Hyuk); ADELI, Hojjat; PARK, Namje; WOUNGANG, Isaac (Eds.). **Mobile, Ubiquitous, and Intelligent Computing: MUSIC 2013**. Berlin, Heidelberg: Springer Berlin Heidelberg, 2014. P. 239–245. ISBN 978-3-642-40675-1.

XU, Yunhong; GUO, Xitong; HAO, Jinxing; MA, Jian; LAU, Raymond Y.K.; XU, Wei. Combining social network and semantic concept analysis for personalized academic researcher recommendation. **Decision Support Systems**, v. 54, n. 1, p. 564–573, 2012. ISSN 0167-9236.

YANG, C.; SUN, J.; MA, J.; ZHANG, S.; WANG, G.; HUA, Z. Scientific Collaborator Recommendation in Heterogeneous Bibliographic Networks. In: 2015 48th Hawaii International Conference on System Sciences. [S.l.: s.n.], Jan. 2015. P. 552–561.

YANG, Chen; MA, Jian; SILVA, Thushari; LIU, Xiaoyan; HUA, Zhongsheng. A multilevel information mining approach for expert recommendation in online scientific communities. **Computer Journal**, v. 58, n. 9, p. 1921–1936, 2014. ISSN 14602067.

YANG, Kun-Woo; HUH, Soon-Young. Automatic expert identification using a text categorization technique in knowledge management systems. **Expert Systems with Applications**, v. 34, n. 2, p. 1445–1455, 2008. ISSN 0957-4174.

YANG, Zaihan; HONG, Liangjie; DAVISON, Brian D. Academic Network Analysis: A Joint Topic Modeling Approach. In: PROCEEDINGS of the 2013 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining. Niagara, Ontario, Canada: ACM, 2013. (ASONAM '13), p. 324–333.

ZEHNALOVA, S.; HORAK, Z.; KUDELKA, M.; SNASEL, V. Evolution of Author's Topic in Authorship Network. In: 2012 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining. [S.l.: s.n.], Aug. 2012. P. 1207–1210.

ZHAN, Zhenjiang; YANG, Lichun; BAO, Shenghua; HAN, Dingyi; SU, Zhong; YU, Yong. Finding Appropriate Experts for Collaboration. In: PROCEEDINGS of the 12th International Conference on Web-age Information Management. Wuhan, China: Springer-Verlag, 2011. (WAIM'11), p. 327–339.

ZHAO, Zhou; YANG, Qifan; CAI, Deng; HE, Xiaofei; ZHUANG, Yueting. Expert Finding for Community-based Question Answering via Ranking Metric Network Learning. In: PROCEEDINGS of the Twenty-Fifth International Joint Conference on Artificial Intelligence. New York, New York, USA: AAAI Press, 2016. (IJCAI'16), p. 3000–3006.

ZHENG, Guoqing; GUO, Jinwen; YANG, Lichun; XU, Shengliang; BAO, Shenghua; SU, Zhong; HAN, Dingyi; YU, Yong. Mining Topics on Participations for Community Discovery. ACM, Beijing, China, p. 445–454, 2011.

ZHOU, Xing; DING, Lixin; LI, Zhaokui; WAN, Runze. Collaborator recommendation in heterogeneous bibliographic networks using random walks. **Information Retrieval Journal**, p. 1–21, 2017. ISSN 1573-7659.

ZHU, Hengshu; CHEN, Enhong; XIONG, Hui; CAO, Huanhuan; TIAN, Jilei. Ranking user authority with relevant knowledge categories for expert finding. **World Wide Web**, v. 17, n. 5, p. 1081–1107, 2014. ISSN 1573-1413.

ZHU, J.; GONCALVES, A. L.; UREN, V. S.; MOTTA, E.; PACHECO, R. Mining Web data for competency management. In: THE 2005 IEEE/WIC/ACM International Conference on Web Intelligence (WI'05). [S.l.: s.n.], Sept. 2005. P. 94–100.

ZHU, Jianhan; HUANG, Xiangji; SONG, Dawei; RüGER, Stefan. Integrating Multiple Document Features in Language Models for Expert Finding. **Knowl. Inf. Syst.**, Springer-Verlag New York, Inc., New York, NY, USA, v. 23, n. 1, p. 29–54, Apr. 2010. ISSN 0219-1377.