A visual UML-based conceptual model of information-seeking by computer science researchers

Cristian Moral *, Angelica De Antonio, Xavier Ferre

Escuela Técnica Superior de Ingenieros Informáticos, Universidad Politécnica de Madrid, Campus de Montegancedo s/n, 28660 Boadilla del Monte, Madrid, Spain

A B S T R A C T

The information-seeking process carried out by researchers is complex and full of differ- ent variables. We have represented this complexity for computer science researchers in the form of a conceptual model. The model is presented in a visual form using the UML mod- eling language, since it allows conveying all the complexity present in such a process with greater clarity. It has been obtained after carrying out two qualitative studies —a focus group and semi-structured interviews— with computer science researchers. The proposed overall model is composed of 4 sub-models: of the documents used in the process, of the tasks undertaken, of the user, and of the information-seeking process context. The concep- tual model proposed can serve for the purpose of better understanding the information- seeking process, for example for librarians or for software designers wanting to provide a support to such task. It can also be useful as a framework to characterize different software solutions aimed to information-seeking in research activities, and to compare them.

1. Introduction

We live in a digital era where it seems easy to find any piece of information in the Internet, where millions of people and organizations produce, share and consume tons of data every second. However, the existence of this huge amount of information greatly hinders the finding of the information that is actually needed. It is difficult to identify which information is reliable and which one is erroneous, incomplete or even false.

Research tasks have benefited from this great availability of information in the last 20 years, but the problem of identifying relevant information with the required level of quality has worsened. Information-seeking is of high relevance for research activities, since the quality of the found information directly affects the quality of the overall research activity. There is a certain degree of variability between researchers regarding the specific information-seeking strategies, but there are important commonalities that can be gathered and are relevant for supporting the task of researchers. Understanding all the elements that form part of the information-seeking process and how they interrelate will provide a deeper understanding of the tasks involved and the challenges for the researcher. Conceptual models are a way of representing elements in a particular domain and their relations, and they can serve as a reference material used to train novices in the field, or to provide support to the tasks represented in different ways, either by means of technology or via specialized professionals working in the field.

* Corresponding author.
E-mail addresses: cmoral@fi.upm.es (C. Moral), angelica@fi.upm.es (A. De Antonio), xavier@fi.upm.es (X. Ferre).

A pictorial conceptual model (Järvelin & Wilson, 2003) provides the information in a conceptual model in a visual form. We believe this kind of representation can serve to more clearly transmit the relevant information to stakeholders due to the excellent human capacity to process visual information. Conceptualizing in this way all the information related to the information-seeking process, and more specifically to the documents used in research, can be very useful for librarianship researchers and staff, as this kind of models make explicit and visual all the concepts and relations they have to manage, in a self-explanatory way.

Conceptual models can also serve as a tool for software designers, guiding the design process along the user view of the relevant elements to consider. In this direction, Preece, Rogers & Sharp state that *"most interface applications are actually based on well-established conceptual models"* (Preece, Rogers, & Sharp, 2015).

Each research field may have differences in terms of the information-seeking process, but we expect that most elements will be shared between researchers of different fields. We have focused in particular in the information-seeking activities carried out by researchers in Computer Science (CS). For the elaboration of a conceptual model about such activities, we have carried out two qualitative studies to be able to discover all the relevant elements for the CS researchers involved in the studies, without any assumptions on our part. This approach has been widely used by many other authors (Ellis, 1989a; Ellis & Haugan, 1997; Foster, 2004; George et al., 2006; Gralewska-Vickery, 1976; Hertzum & Pejtersen, 2000; Makri, Blandford, & Cox, 2008; Meho & Tibbo, 2003). To be more specific, we chose to apply a focus group and a series of in-depth semi-structured interviews. Such techniques are described in Krueger and Casey (2015) and DiCicco-Bloom and Crabtree (2006), respectively. As a result of the qualitative studies, we obtained a set of 169 codes categorized in 13 categories. The codes have formed the basis for creating the conceptual model, structured as a high-level general model, further detailed through four sub-models, namely a conceptual sub-model of the documents used during the information-seeking activities, a conceptual sub-model of the tasks that have to be carried out when seeking information, a third sub-model of the user – in our case, a researcher in the CS field –, and, finally, a sub-model conceptualizing the context in which an information-seeking activity is carried out. As far as the authors know, there are no detailed conceptual models about the information-seeking process represented through visual means, like the one proposed in the present work.

In the following section we review the literature in order check what other authors have proposed regarding the modeling of the information-seeking process. In Section 3 we detail the empirical design of the qualitative studies and how we have formalized the representation of the conceptual model using a well-known and widely-used standard representation (UML). Subsequently, Section 4 introduces the main model representing all the actors and components that can intervene in an information-seeking process. From there, each section is used to explain each one of these entities more in detail with the help of a more concrete sub-model: Section 5 details the characteristics of the documents, both individually and as a collection; Section 6 presents the tasks that are performed in an information-seeking activity and how they are carried out; Section 7 details which are the main characteristics and specificities of CS researchers when seeking information; and finally, the context in which an information-seeking activity takes place is presented in Section 8. In Section 9 are presented some of the practical applications of the conceptual model presented in this article. Then, in Section 10 is discussed where are our results situated in comparison with the most relevant related works, and Section 11 raises the possible limitations of the proposed model. Finally, Section 12 summarizes the main gathered conclusions and defines the future lines of work based on these results.

2. Literature review

We can find many authors in the literature who have tried to model, from different points of view, the information-seeking process undertaken by users of different profiles, working in different contexts and/or in different fields (Case, 2012; Ford, 2015; Wilson, 1999a). Bates (1989) stated that the typical information retrieval model based on the best-match approach is not valid, as fulfilling an information need is a complex activity that requires more than a query. The need evolves during the process and the context of use greatly affects how the process is performed. Bates concludes that each search provides the user with a small chunk of the information he/she needs, and that the user uses different search strategies and different sources of information in order to fulfill his/her information-seeking purpose. Then, it seems essential to understand both the user and the tasks he/she carries out – as well as the context in which he/she carries them out – in order to properly model the information-seeking process.

Many authors have derived information-seeking models detailing which are the stages through which a user usually passes during an information-seeking activity. One of the most relevant and referred models, proposed by Ellis (1989a), firstly consisted of 6 stages – "*Starting*", "*Chaining*", "*Browsing*", "*Differentiating*", "*Monitoring*", and "*Extracting*" – and was derived from the study of academic researchers in social sciences like psychology, sociology, politics or geography. However, the model has been also validated in other fields like Physics, Chemistry, and English Literature (Ellis, 1993; Ellis, Cox, & Hall, 1993). Later, Ellis updated his model with two new stages – "*Filtering*" and "*Ending*" – (Ellis & Haugan, 1997) after studying the behavior of professionals – non-academic researchers – and research scientists in an industrial environment – an oil research center. More recently, Meho and Tibbo (2003) proposed to complete the Ellis' initial model with four new stages – "*Accessing*", "*Networking*", "*Verifying*," and "*Information Management*" – after analyzing a set of structured and semi-structured interviews performed on sixty social sciences researchers. Later on, Makri et al. (2008) modeled the information-seeking behavior using 14 low-level stages – among which we can find most of those proposed by Ellis –

grouped into three high-level stages — "*Identifying and locating*", "*Accessing*", and "*Selecting and Processing*". In this case, the study targeted academic lawyers.

There are many other authors that, instead of defining the process only as a set of tasks that are performed by the user, they also include the stages through which a user passes during the information-seeking process. This is usually called information behavior. As an example, Marchionini and White (1997) provide a behavioral model of the information-seeking process by defining a set of progressive and iterative tasks — "*Problem formulation*", "*Express the information need*", "*Examination of results*", "*Reformulation*", and "*Use of the found information*" — together with a set of stages that are more related to the user — "*Recognize a need*" and "*Accept a challenge*". In the same direction, Foster (2004) identifies a set of tasks that he calls core processes — like "*Breadth exploration*", "*Keyword searching*", "*Browsing*", "*Chaining*", "*Reviewing*", "*Identifying keywords*", "*Refining*" or "*Verifying*" — and categorizes them into three main activities —"*Opening*", "*Orientation*", and "*Consolidation*". Foster finds that these tasks and activities take place in three levels of contextual interaction: the "*Cognitive Approach*" level, where the user processes the information and takes decisions; the "*Internal Context*" is related with the user's feelings and thoughts; and the "*External Context*", that deals with the social aspects involving the user while seeking information. Finally, Kuhlthau (1991) also models the information-seeking process as a sequence of six user-centered stages: "*Initiation*", "*Selection*", "*Exploration*", "*Formulation*", "*Collection*", and "*Presentation*". Her model, which is one of the most relevant and cited in the literature, focuses in trying to understand the user's behavior while seeking information from a psychological and sociological point of view.

Even if many of the models that can be found in the literature have been explained using a more or less extensive and more or less detailed verbal description — like for example those presented above —, we also find some authors that have preferred to also use a pictorial representation to present their information-seeking models. One of the best representatives of this approach is the pictorial model of the user's behavior while seeking information proposed by Wilson (1981), as he summarizes the process using a graphical diagram where the main actions and objects are represented. Some years later, Wilson expanded his model including other external aspects — "*Economical*", "*Sociological*", "*Situational*", and about the "*Information sources*" — that can also intervene in the information-seeking process (Wilson, 1997). Later on, Niedźwiedzka (2003) proposed a new information behavior model based Wilson's revised model. In her model, Niedźwiedzka maintained all the components proposed by Wilson, but she reorganized and refactorized them, especially the Wilson's intervening variables — "*Personal*", "*Role-related*", and "*Environmental*" — that in her model are part of the context, which in turn affects all the process stages. Additionally, Niedźwiedzka's model reflects that the process can be activated at any of its stages by different mechanisms, and not only at the beginning. Leckie, Pettigrew, and Sylvain (1996) derived their graphical information-seeking behavior model for researchers in general — even if the study focused on engineers, health care professionals and lawyers, they claim that the model can be generalized. In their case, an information-seeking process starts at specific stage — "*Work Roles*" — and, after performing a set of intermediary stages, it ends up with the "*Outcomes*" stage.

To the best of our knowledge, among the existing pictorial models of the information-seeking process, only those from Krikelas (1982) and Benardou, Constantopoulos, Dallas, and Gavrilis (2010) have been presented as conceptual models. In both cases, even if the representation is not framed in a standard representation like UML, their approach is very similar, and they visually illustrate concepts and relationships in a similar way. In the first case, Krikelas describes the main high-level concepts and actions related to the information-seeking behavior — like "*Information needs*", "*Sources of information*", "*Information gathering*" and "*Information giving*" — and then illustrates them in a pictorial conceptual model relating them in terms of association — for example, "*information gathering*" occurs when the user's "*information need*" does not have to be satisfied immediately and can be deferred — or of generalization — a "*source preference*" can be of two types, internal or external. On the other hand, Benardou proposes a pictorial conceptual model containing very high-level concepts, like "*Actor*", "*Information object*", or "*Method*" that are related through a set of tagged relationships —as done in the association relationships in UML. In this case, moreover, the model aims at conceptualizing the overall research activities performed by scholars, including a lot of other sub-activities apart from seeking information, like, for example, writing a paper.

Focusing on the area we have studied, to the best our knowledge, no models related to information-seeking performed by CS researchers have been proposed yet. Some authors have separately studied how engineers (Freund, Toms, & Waterhouse, 2006; Hertzum & Pejtersen, 2000; Kwasitsu, 2003; Leckie et al., 1996; Holland & Powell, 1995; Pinelli, 1991) or academic or industrial research scientists from different fields (Ellis, 1989b, 1993; Ellis et al., 1993; Foster, 2004; Meho & Tibbo, 2003) seek information, but none of them has focused on researchers with a strong technical education background, especially in information and communication technologies, like CS engineers. Ellis and Haugan (1997) studied how engineers and research scientists perform their information-seeking activities in an industrial environment, but only one of the 26 participants of the study was a scientist and had a technical education background — mechanical engineering. Similarly, Anderson, Glassman, McAfee, & Pinelli (2001) also studied how 872 aerospace engineers and scientists looked for information. As far as we know, only O'Brien and Buckley (2005) have to some extent studied the information-seeking activities when performed by users with a strong background in ICT —but in this case they are not researchers —, as they propose an information-seeking behavior model for programmers involved in software maintenance. This model is based on some of the models mentioned above, but has been adapted to the specific needs existing in the studied context —for example, seeking information is presented as an iterative process, or the stage leading to the identification of the problem is omitted as in the software maintenance the context of the problem is usually defined in advance by somebody. At the end, this non-linear model proposes five stages. The user has to first be "*Aware of the Problem*", which requires an initial understanding of it. Then, the information seeker has to "*Focus the Formulation*" and narrow the problem to formulate specific queries. At this point the user has to start the "*Information Collection*" stage where he/she tries to identify what sources can be relevant, which usually implies performing chained searches based on who refers to these sources — forward snowballing —, or who is referred by in the sources — backward snowballing. Once relevant sources have been identified their content is browsed in order to differentiate which information is potentially relevant and has to be extracted, and which one seems to be non-relevant and then has to be omitted. After this, the seeker has to "*Examine these Results*" more in detail to determine if they are actually relevant to fulfill his/her information need. Finally, the process ends when the user considers that his/her information need has been satisfied — "*Problem Solution*". Shortly after, Buckley, O'Brien, and Power (2006), after empirically evaluating their model with two maintenance programmers, extended their model with a sixth stage, "*Information Prompted Action*", between "*Examine Results*" and "*Problem Solution*" in order to accommodate the actions allowing to provide directional information that can be used as a base to facilitate the performance of further information-seeking activities.

Some authors have proposed theoretical information-seeking or information behavior models, that are meant to be general (Bates, 1989; Godbold, 2006; Leckie et al., 1996; Marchionini, 1997; Wilson, 1997). Our premise, however, is that the information-seeking process depends on many factors, like the user — in our case a CS researcher —, his/her purpose, the context in which the seeking is carried out, and the tasks that are performed during the process to reach the mentioned purpose. Cheuk and Dervin (1999) reached the same conclusion after analyzing how auditors, engineers and architects perform their information-seeking activities using the "*Sense-Making Methodology*" (Dervin, 2003, 1992; Dervin & Nilan, 1986). The analysis of the study, based on the identification of three aspects — the "*Contextual time-space situation*" in which the process is performed; the information need or "*Gap*" to be covered; and how the user makes "*Use of the obtained information to construct new knowledge*" — resulted in the identification of ten different contexts — called "*Information-seeking situations*". Seven of these situations — "*Task initiating*", "*Focus forming*", "*Idea assuming*", "*Idea rejecting*", "*Idea confirming*", "*Idea finalizing*", and "*Idea sharing*" — were commonly experienced by the three types of participants, while the three others —"*Approval granting*", "*Design generating*", and "*Approval seeking*" — were primarily experienced only by architects. Then, they concluded that, even if there exist many commonalities between different workplace domains, there also exist specificities that can affect the information-seeking process. Byström and Järvelin (1995) also propose an information-seeking model that reflects the dependence of the information-seeking process on the researcher characteristics and preferences — called "*Personal factors*" and "*Personal style of seeking*" — and on the context in which the seeking is being carried out — called "*Situational factors*". Additionally, they classify the tasks into five categories based on their determinability, which is closely related with their levels of uncertainty and complexity. As filling complex information needs usually results in the performance of several activities to iteratively reduce the complexity and uncertainty of the problem, the model is structured as a feedback loop so that the result of each activity is used as input in the following activity. Similarly, Domik and Gutkauf (1994) stated that, in order to improve the information-seeking process performed through information visualization systems, there are four factors that have to be taken into account. "*User model*" contains the information owned by the system about a specific user in terms of demographic data, experience, abilities, or limitations, among others; "*Problem domain/task model*" defines which are the best pictorial representations based on the domain in which the visualization is being performed; "*Resource model*" concentrates the knowledge related to the software and hardware being used to carry out the visualization; and "*Data model*" describes the main characteristics of the data to be visualized. Even if information visualization and information-seeking are not exactly the same activities, they are very closely related, and are often performed together, and then the aspects intervening in one of them can be useful to better define the other one.

3. Materials and methods

3.1. Formalism used in the conceptual model

A conceptual model is a tool that allows describing the most important aspects, objects or actions that are involved in a given situation. These aspects are called concepts. Besides, a conceptual model provides information about the relationship between the different concepts. Each of these relationships, which can involve two or more concepts, is specific and defines how the concepts relate to each other. In software engineering, for example, conceptual modeling is used to provide an abstract description of the main aspects that have to be considered to make up good functional and non-functional requirements leading to the design of an effective and efficient system (Boman, Bubenko, Johannesson, & Wangler, 1997).

There are two main approaches to represent conceptual models: using textual representation, or pictorial representation. Commonly, as explained in Section II, the first option has been the predominant approach in the information behavior research field due to its flexibility. However, reading and understanding a complex and extensive model described through text can be tedious and counterproductive. Additionally, "*the flexibility of textual representation may lead to an ambiguous description of the [...] model*" (Onggo, 2010, p. 340) and, if the conceptual model is used as a communication means between the stakeholders, a textual representation is useful "*only if all stakeholders understand the language used in the texts*" (Onggo, 2010, p. 341). Pictorial conceptual models, nonetheless, allows depicting a great amount of concepts and relationships in a condensed and more comprehensive representation (Onggo, 2010, p. 345).

Modeling the information-seeking process through a pictorial conceptual model will allow us to clearly illustrate the data obtained from the qualitative studies in a compact and non-ambiguous way. Additionally, pictorial conceptual models are a powerful tool that provides us with high flexibility, as they can be easily extended if needed. In any case, we consider that

Table 1
UML elements used in the conceptual model.

Name	Representation	Description	Example
Class	CLASS A	Class A represents a concept.	*Researcher* is a concept. RESEARCHER
Generalization	CLASS A; CLASS A_1, CLASS A_2	Generalization is a taxonomic relationship between a more general element (A) and other specific ones (A1 & A2).	*Filtering Task* and *Reading Task* are two kinds of *Information-Seeking Tasks.* INFORMATION-SEEKING TASK; FILTERING TASK, READING TASK
Composition	CLASS A; CLASS B	An instance of class A is composed by one or more instances of class B.	A *Document Collection* is composed by at least one *Document.* DOCUMENT COLLECTION; DOCUMENT
Association	CLASS A; relationship; CLASS B	Class A is related to Class B as defined by “relationship”.	An *Information-Seeking Activity tries to achieve* an *Information-Seeking Purpose.* INFORMATION-SEEKING ACTIVITY; tries to achieve; INFORMATION-SEEKING PURPOSE

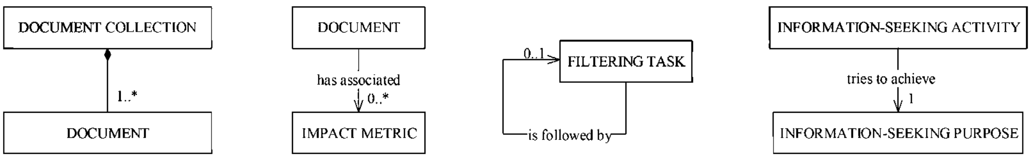

Fig. 1. Some examples on the use of UML cardinalities.

textually explaining the model is still very relevant, and then we also provide a textual explanation. However, we believe that having a comprehensive and non-ambiguous pictorial representation perfectly complements and improves this textual explanation, especially when the reader requires analyzing a very specific — and probably small — part of the model, or when he/she requires getting a general overview of it.

In this paper, we have used part of the notation for class diagrams proposed by the Unified Modeling Language (UML) (Object Management Group, 2016) to construct our pictorial conceptual model, as it is a well-known and standardized modeling language that is widely used mainly in software engineering (Stevens & Pooley, 2006), but also in other fields like design of simulation models (Richter & Marz, 2000). By doing this, we provide the model with a high degree of understandability and unambiguity as the notation, its use and its meaning is very well defined. Table 1 summarizes and briefly explains the notation used in the model, with examples extracted from our contribution.

Both the composition and association relationships between concepts can also have a cardinality attached that states how many instances — minimum and maximum — of the target concept can be related with one instance of the origin concept. Fig. 1 illustrates some examples of cardinalities: a document collection is formed by at least one document —* represents an undefined number, and then it indicates that there is not a defined maximum number of documents that can form a document collection —; a document may not have an impact metric associated, but it can also have an undefined number

Table 2
Main characteristics of the participants of the qualitative study.

	Focus group									Interviews							
	P1	P2	P3	P4	P5	P6	P7	P8	P9	P1	P2	P3	P4	P5	P6	P7	P8
Research experience (in years)	4	3	1	2	2	15	25	20	27	2	6	4	17	31	37	28	15
Gender	F	M	F	M	M	M	M	M	F	F	M	F	M	F	M	F	F
Age	36	31	43	27	44	40	49	45	53	32	29	43	40	49	65	50	41
Job title	Ph.D. student					University professor & researcher				Ph.D. student			University professor & researcher				

of them; a filtering task can be followed by another filtering task or by none; and an information-seeking activity always usually tries to achieve a specific information-seeking purpose.

3.2. Qualitative studies: empirical design

As mentioned above, our research specifically focuses on the information-seeking process in the CS field, and that's why we recruited all the participants in both qualitative studies — focus group and interviews — between researchers in CS. Apart from this, we did not establish any specific selection criteria, although we tried to have a sample with more or less the same number of males and females, trying also to have an as wide as possible spectrum of age covered by the participants — concretely, they were between 27 and 65 years old. Additionally, we also aimed at having participants with different research profiles and background, and that's why we decided to have both many experienced and novel CS researchers, as we believed that this could be a good way to evaluate if experience influences the way the information-seeking process is performed, and if so, how it affects the process. As the aim of our qualitative studies was not to identify the most common profiles of CS researchers based on statistics, but only to discover relevant information about how information-seeking activities are undertaken by computer scientists, we selected nine CS researchers for the focus group, while eight different CS researchers were individually interviewed. Table 2 summarizes the main characteristics of the CS researchers that participated in the qualitative studies.

The size of the sample can be considered sufficient in the context of a qualitative study, where completeness is attained when theoretical saturation is achieved, that is when collecting and analyzing new data do not result in generating new knowledge. In our case, the codification of the first interview generated 88 of the 189 codes obtained in the first phase of the coding, before the three validation iterations that were later performed. From there, the discovery of new codes decreased from interview to interview: 36, 22, 21, 17, 2 and 3 new codes appeared after coding interviews 2–7. Finally, in the eighth interview no new codes were obtained. Then, due to the decreasing evolution of the appearance of new codes and the no appearance of new codes in the codification of the eighth interview, we considered that theoretical saturation was achieved and then we stopped performing new interviews to collect data.

At that point, the volume of data collected was considerably large, as 103 min of audio were recorded in the case of the focus group, while the eight interviews produced around 450 min of audio recording. The transcription of the answers given by the participants in both studies resulted in around 40 pages of free-text, containing around 15,000 words, which required several codification iterations and validation phases that took place over several weeks. With respect to the questions that led both studies, in the case of the focus group, only a few open-questions were posed to contextualize the session, while in the interviews 18 questions were stated, most of them also open-questions. In both cases, questions were only used as a guide helping the moderator and interviewer to properly conduct the sessions, as the answers from participants and the evolution of their discourse and behavior was the main thread that guided the sessions.

Then, all the data collected from both qualitative techniques were analyzed using an inductive approach — Grounded Theory (Corbin & Strauss, 2014; Glaser & Strauss, 2009) — allowing to generate a theory from the analysis of a set of raw data extracted from real CS researchers without using —at least not before the theory statement — any previous knowledge or reference framework. This approach has also been used by Ellis in his studies (Ellis, 1989a, 1993; Ellis & Haugan, 1997; Ellis et al., 1993).

Our approach mostly matches the theoretical sampling used in classic Grounded Theory (Glaser, 1978), where an open-coding of the data collected is initially performed (Saldaña, 2015). From these initial codes emerged a set of more abstract concepts allowing to relate similar codes that were later categorized. These three steps were iteratively repeated until the obtained codification was stable and the coder considered that all the answers of the interviewees were coded correctly. Additionally, to validate the resulting concepts and categories and their use when coding the transcriptions, the codification process was also performed by two additional coders. None of them were involved in the data collection. In order to code the transcribed interviews, both coders were provided with the set of codes generated by the first coder, in order to use them as a framework. Once both of them had coded the transcriptions, a new refinement process was performed to select only those codes that were used by at least two of the three coders to tag the same fragment of text. This comparison also allowed us to identify and discard irrelevant and ambiguous codes, and to include new concepts that were not present in the initial set of codes. All the codes resulting from this refinement were included in the coding system, provided they were used at least one time to code the transcriptions of the raw data.

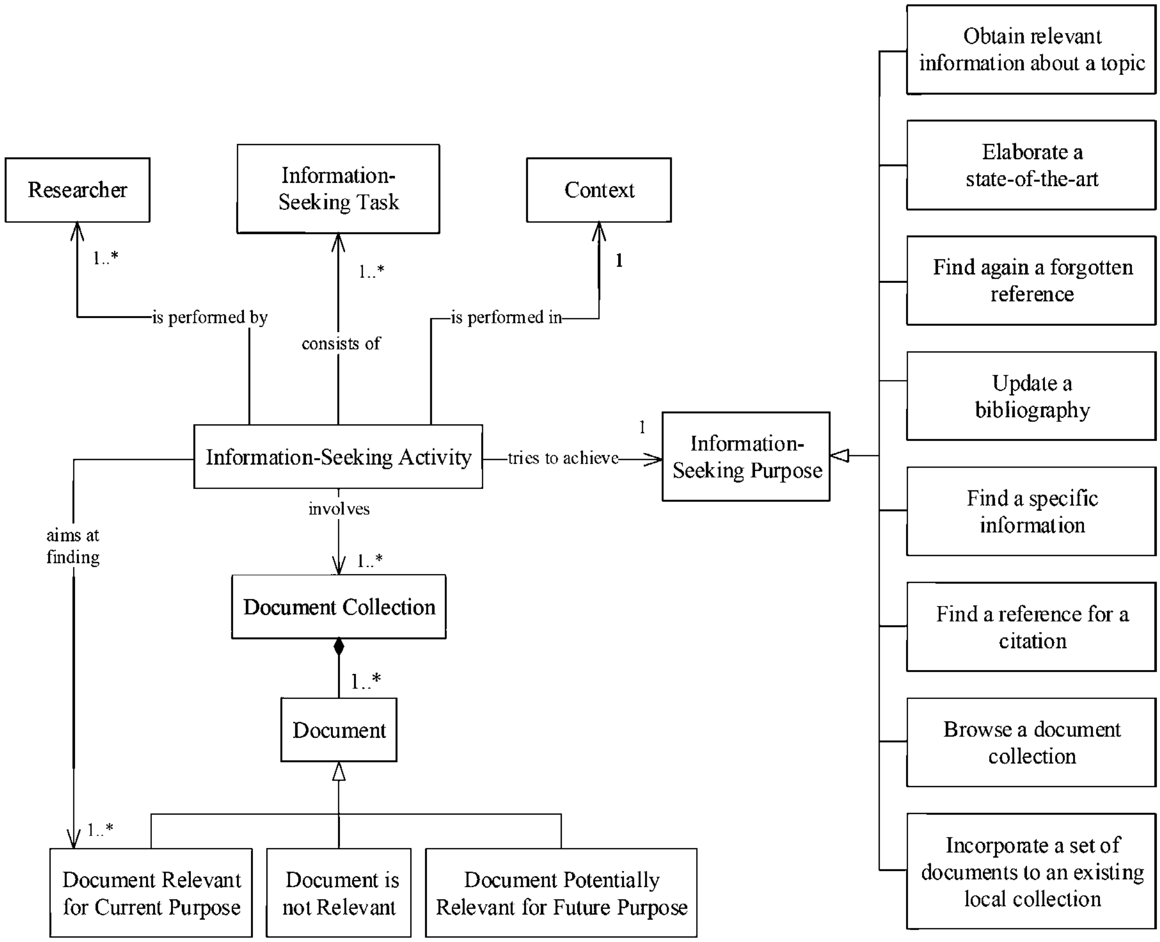

Fig. 2. Main concepts intervening in an information-seeking process.

As a result, 169 codes emerged from this inductive coding process. These codes were later grouped into 13 categories: "*Archiving*", "*Search*", "*Corpus size*", "*User wishes*", "*Workspace*", "*User tasks*", "*Identified difficulties*", "*Information-seeking process*", "*Meta-information*", "*Reference manager*", "*Reading process*", "*User profile*", and "*Collaborative work*". More details of the codification and validation processes, the obtained results and considerations about their validity can be found in Moral, de Antonio, Ferre, and Lara (2015).

We have considered all the emerged concepts except the ones exclusively related to the documents storage and management task, or to the visualization of and interaction with document collections in a digital system. We chose not to consider those elements for two reasons: first, because both aspects are complex enough to be treated and investigated separately; and second, because most of the obtained knowledge — especially in the case of visualization and interaction — were rather assumptions than real facts from the observation of the current context of use. This is due to the fact that almost none of the participants and interviewees made use of any specific tool for storing, managing, visualizing and interacting with research document collections. Then, all these tasks, even though they are as relevant as the ones included in this model, will be tackled in future works.

Nonetheless, the goal of performing a grounded theory approach is to generate a theory, not only a categorized codification system. This is precisely what we aim to do in this paper, as the conceptual model we propose allows not only to summarize all the knowledge emerged from the qualitative studies in form of codes, categories and memos, but also to formulate a theory of which are the main aspects, actors, and relationships that, according to the participants and interviewees, allow to conceptually describe the information-seeking process in the CS domain.

4. Information-seeking model

Fig. 2 illustrates the core entity of an information-seeking process, the information-seeking activity, that, in turn, can be defined by the information-seeking tasks — at least one — that are performed by one or more CS researchers in a given context in order to find the relevant document(s) included in one or more collections that allow him/her to achieve a given information-seeking purpose.

In overview, the model presents information-seeking as a motive-driven activity, that can be decomposed, according to Marsh (2003), as follows: an "*activity*" is defined by its "*goal*" — which in our model is represented through an information-seeking activity that allows to fulfill an information-seeking purpose —; at a lower level, each of these activities is achieved by performing a set of "*actions*" — in our case information-seeking tasks — in order to meet "*specific goals*" leading to the achievement of the activity purpose; finally Marsh claims that actions are composed of a combination of "*operations*", that are well learned by the user and then are *"unconsciously performed processes [...] that are performed without conscious thought or effort"* (Marsh, 2003, p. 87). In our case, these operations are not explicitly represented as concepts, but nevertheless its recognition and understanding is the main objective of the analysis of the context of use observed in the qualitative studies — and then they are the basis of the conceptual model —, as they are essential to really understand what the real behavioral, conceptual and operational problems are when seeking information. We have decided to adopt this decomposition as the information-seeking purposes we present are rather high-level objectives whose complexity depends on a lot of aspects like the CS researcher himself/herself, the context, or the topic being investigated. Nevertheless, the aim of our information-seeking tasks — explained in Section 6 — is very well defined: a filtering task aims at discarding or selecting a set of documents from a bigger collection that matches a set of criteria; an exploration task allows the CS researcher browsing more in detail a set of documents to determine which of them are actually relevant or to identify some relevant information; and the aim of a reading task is to deeply enter into the content of a document. Then, it makes more sense relating the final information-seeking purpose with an information-seeking activity formed by a set of information-seeking tasks whose performance is interleaved — and then provides more flexibility —, than to the tasks themselves.

Most of the literature related to the modeling of the information-seeking behavior defines the existence of an information need as the process trigger. However, as Wilson stated when he presented his first model (Wilson, 1981), using the term information need could lead to confusion as any human need can be divided into three interrelated features, namely physiological, affective, and cognitive aspects. Case (2012) also highlights the difficulty of defining what is a need in general, and more specifically an information need, as users can have many motivations to carry out an information-seeking activity, like seeking answers, reducing uncertainty and/or making sense of a situation. However, regardless the meaning assigned to information need, it is assumed that an information-seeking activity starts when the user becomes aware that he/she has less knowledge than the one he/she needs to solve a problem (Krikelas, 1982). In our approach, instead of making the information-seeking process revolve around this information need, it is led by the resolution of the problem, that is the purpose that wants to be achieved. By doing this, we allow our model to represent information-seeking as an adaptive process, as it depends, at least, on the specific goal of the CS researcher.

Due to the diversity of information-seeking activities, a CS researcher may have to achieve a lot of different purposes. Nonetheless, as mentioned before, in our model we have considered only those that have been mentioned by the interviewees during our study (Moral et al., 2015), as they are probably those that are more relevant and usual for them. It is important to differentiate between these purposes because, in fact, each of them has different objectives and specificities that are implicitly attached to them. In order to support concrete information-seeking activities it is important to make the main purposes of a CS researcher explicit:

(a) *Obtain relevant information about a topic* refers to the commonly initial exploratory phase when the CS researcher needs to delve deeper into a topic. This usually means that he/she lacks or needs to increase his/her relevant knowledge related to the topic, like its specific terminology, which are the most relevant publication venues that regularly address it, or which are the most relevant authors who produce relevant scientific documents in the topic. In summary, the CS researcher wants to build a mental model including the main concepts within a topic.
(b) *Elaborate a state-of-the-art* means trying to find all the relevant documents that are related to a topic in order to draw its current situation in the research field. In this case, the information-seeking activity has to be exhaustive and systematic in order to ensure the completeness of the obtained information.
(c) *Find again a forgotten reference* corresponds to the need of finding the bibliographic reference to a document that the CS researcher has already read — or at least explored — before, but without remembering enough specific data — like its title, its author(s) or its publication venue — that would allow him/her to directly find it again.
(d) *Update a bibliography* refers to the need of finding documents related to a topic that have been published recently — or from a given date — in order to include updated bibliographic references in a new document to be published.
(e) *Find a specific information* deals with the need of obtaining a very concrete information, like a definition or a formula. It is the most directed and narrowed information-seeking purpose as the kind of output to be obtained is previously known by the CS researcher, and then it is easier for him/her to determine if the result is the expected one or not. Besides, in this case, the CS researcher usually knows — to a greater or lesser extent — how to obtain the desired information — for example where to look for it, which terms to use in the query, or which authors should preferably be considered.
(f) *Find the reference for a citation* also represents the need of finding a specific information, which in this case is the bibliographic reference to a document that supports or contradicts — depending on the aim of the citation — a given statement or hypothesis. Even if this information-seeking purpose is also directed, achieving it is not so straightforward because its formulation is usually not so direct, and then it often requires the CS researcher to analyze more

in detail one or more full documents — or at least part of them — to determine which ones contain the sought information and then can be cited. In this case, besides, the quality of the document to be cited usually also plays an important role in the selection.

(g) *Browse a document collection* has to do with the aim of exploring a set of documents at a general level. In this case, the CS researcher does not seek for a specific kind of information, but, on the contrary, he/she only wants to know more about a set of documents, like for example which are the main topics they address.

(h) *Incorporate a set of documents to an existing local collection* characterizes the desire of a CS researcher to evaluate a set of documents in order to identify which of them should be kept and added to his/her local collection of documents. In this case, documents are not selected to fill a specific information need, but only to update and thicken the CS researcher's local collection. As an example, when a CS researcher receives a new issue of a research journal, but due to lack of time he/she is not able to read all the articles included in it, he/she may just want to identify which ones might be relevant or interesting for him/her and mark them somehow to make them easily accessible in the future. For example, storing them in a specific folder of his/her computer, named "To be read".

While seeking information, CS researchers can find documents that are not relevant for his/her current purpose but that can be potentially relevant for future purposes as their contribution is novel and is related, at some level, with his/her topics of interest. Other times, a CS researcher can determine that a given document is not relevant for his/her current purpose, and that there is little chance for it to become relevant to achieve future information-seeking purposes. Identifying this last case can be very useful as knowing that a given document is probably not going to be useful in the future can make the information system avoid showing it again in future information-seeking activities, or at least to somehow indicate that it has already been processed and tagged as non-relevant.

In the next section the high-level concepts implied in an information-seeking activity are detailed, namely documents, information-seeking tasks, CS researcher, and context.

5. Document sub-model

Researchers in general, and specifically CS researchers, typically need to handle hundreds or thousands of documents grouped around one or more document collections during their information-seeking activities. Fig. 3 presents the complete modeling of the concepts "*document*" and "*document collection*". To begin with, a document collection can be physical —if it only contains paper documents like printed books or printed articles —, or digital — if it only contains a set of digital documents. In the first case, there are physical libraries — for example in a University— where thousands of books, journals and other documents are made available to the users, but there are also physical local collections that are formed by a set of paper documents that have been locally archived over the time by a CS researcher. Analogously, it is common for CS researchers to have a digital local collection formed by a set of digital documents stored over the time, but there are also digital libraries that are made available by most of the main editorials —for example, Elsevier Science Direct, ACM Digital Library, or IEEE Xplore Digital Library. In fact, most of the interviewees asserted that digital libraries are increasingly becoming the main source used by many CS researchers to seek information, as they contain a huge amount of past and recent publications. Moreover, they allow accessing the documents from anywhere through the Internet and they provide powerful search tools that simplify and cheapen —in terms of time and money— the process of locating and obtaining the desired documents. Finally, specialized forums and websites can also be considered as collections of forum posts and webpages, respectively.

In research, it is also usual to employ style guides that define how to write and cite research documents, but they only cover some of the concepts related to a research document, like which are their sections or which publication details have to appear in a bibliographic reference — some examples are the style guides (APA, 2001; University of Chicago, 2010; Feder, 2006). However, our objective is to provide a more detailed description, including also the relationship between the concepts. To some extent, this has been already made for journal articles in the "*Archiving and Interchange Tag Suite*" project developed by the National Library of Medicine of the USA where 4 different XML-based tag sets are offered to facilitate the archiving, interchange, publication, and authoring of journal articles (NLM, 2012). Among the hundreds of tags that are defined in the model, we can find, for example, "*journal-title*", "*publisher*", "*pub-date*", "*abstract*", or "*fig*". This model, besides defining many other concepts related with a journal article, also reflects some of the relationships that exist between them, like the ones between "*publisher*" and "*publisher name*", "*pub-date*" and "*month*" and "*year*", an author — "*contrib*" — and its affiliation —"*aff*" — or between an article and its meta-information — "*article-meta*" — like its "*article-id*" — for example its DOI — or its "*title-id*".

Apart from this, in our opinion, considering the variety of sources of documents is essential to adequately support the information-seeking activities in the research domain. Any kind of document containing explicitly written information can be considered as a source of information. According to the interviewees, the document types that are more common in research are articles, academic theses, books, technical reports, forum posts and webpages. In turn, articles can be divided into journal articles and conference papers; there are classical text books consisting of a set of chapters, but also books presented as a compendium of documents — typically journal articles or conference papers —; and an academic thesis can

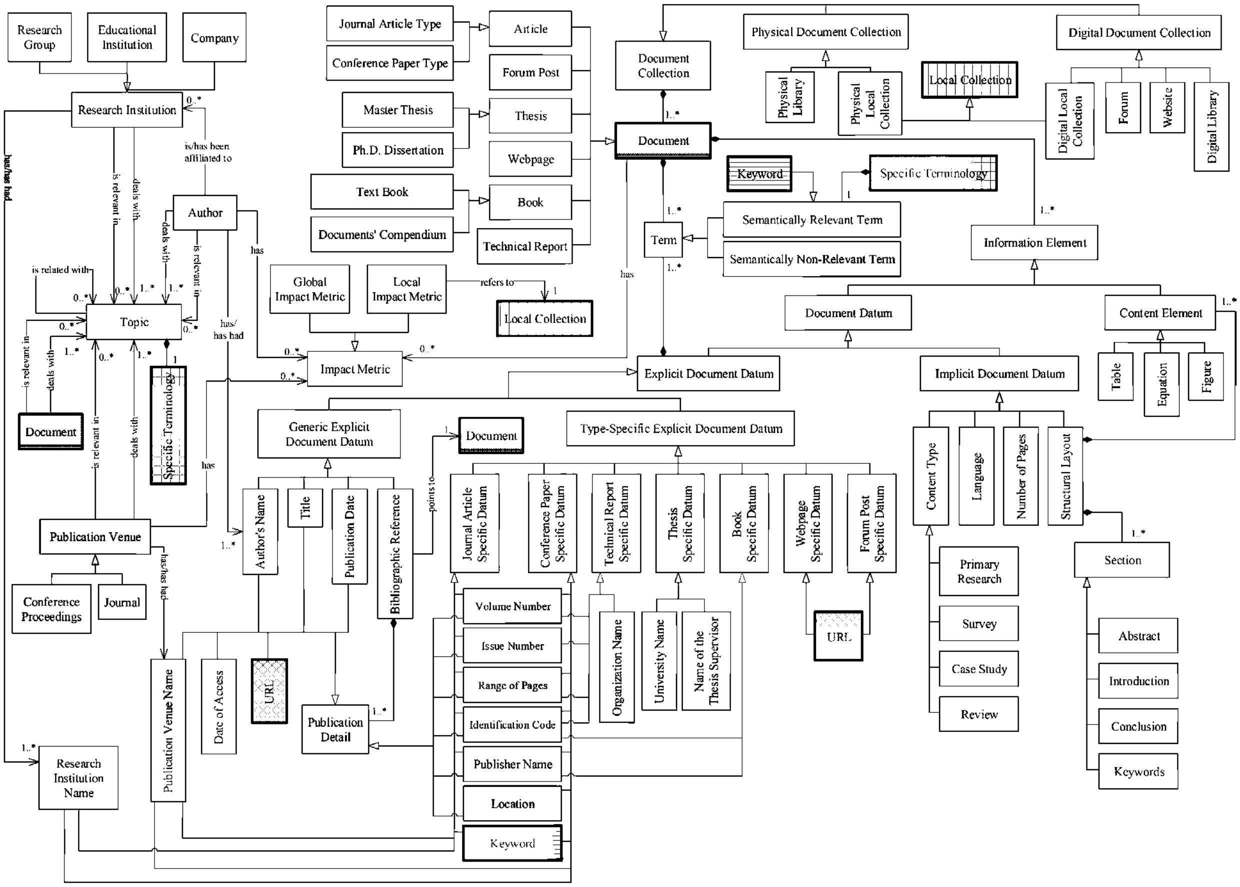

Fig. 3. Document sub-model. Concepts *Document, URL, Keyword, Specific Terminology* and *Local Collection* are duplicated in the sub-model for the sake of clarity. They are represented by a specific pattern (outside the UML standard) to highlight that they stand for the same element.

be a Master thesis — of any kind — or a Ph.D. Dissertation. This subdivision is especially relevant in the case of the articles and theses as, even if their format, structure, and type of content is very similar, interviewees clearly considered them as different types of documents for some reasons, like the difficulty of publication, the novelty of their contribution, and the maturity and reliability of their results in the case of the articles, or the extent of the contribution in the case of the theses. Obviously, these are only some of the types of documents that may be used in the research field, but new ones could be safely added to the sub-model as new types or subtypes of documents, for example, editorial articles, short papers, extended abstracts, posters, position papers, or undergraduate thesis.

At the most basic level, a document consists of a set of terms. These terms, in turn, can be divided into semantically relevant terms — that is those terms that provide relevant information about the topic(s) addressed in the document, like for example those that are provided as keywords or those that are part of the specific terminology of a topic —, and semantically non-relevant terms, which are the rest. At a more conceptual level, a document can be seen as a set of components — that we call information elements — that contain relevant information provided by its content elements like tables, equations, and/or figures, but also some kind of meta-information related to the document and it content. These document data allow to identify the document — who has written it, when it has been written, where it has been published… —, but also to obtain information about its structure — which sections are present, which type of document it is, how long it is… — or about its contribution — which topic(s) it addresses, which is its contribution… —, among others. Some of this information is explicitly provided by some of the terms explicitly written in the document, while some other are implicitly provided, and then have to be inferred from the structure and content of the document.

5.1. Explicit document data

CS researchers usually read and use — to a greater or lesser extent — some of the explicit document data while looking for information for a variety of reasons, as for example to determine the quality of a document. Some of these data are generic and are typically present in any kind of document, while others can only be found in documents of a specific type.

In the first group, we find the name(s) of the author(s), the title and the publication date as primary data, but also a set of bibliographic references, each of them consisting of a subset of data about the document it refers to.

In our model, we propose a flexible way to represent bibliographic references as, instead of predefining which are the fields — in our model called *publication details* — that are standard based on the type of the referenced document, we allow defining each bibliographic reference individually, according to the publication details it contains. This is especially relevant because some supposedly necessary fields are sometimes not present, and sometimes there are additional data that have to be included as publication details. As an example, a bibliographic reference is supposed to always contain a title, one or more authors, and a publication date — among others —, but it can also contain other data that do not appear explicitly nor implicitly anywhere, like the last time the availability of a remote source — like a webpage or an open-access digital journal article — was checked, or the URL from where the referred document has been retrieved.

Beyond generic explicit data, a document can have other data that are specific to its type. In fact, it is common to find explicitly written in an article the name of the publication venue where it has been published — that is the name of the journal, or the name of the conference proceedings —, along with the name of the research institution to which each author is affiliated. In the case of an academic thesis, the explicit data that can be more commonly found are the name(s) of the supervisor(s) and the name of the university where the thesis has been developed. Finally, both webpages and forum posts have as specific data the URL that gives access to them.

In journal articles, conference papers and technical reports it is also possible to find the volume and the range of pages in which the document can be found. As conferences are usually annual, there is only a volume of conference proceedings per year, and then only journal articles and technical reports use to have an issue number associated. Nevertheless, it is not uncommon to see explicitly indicated in a conference paper the location where the conference took place. Finally, these three types of documents and also books usually have an identification code explicitly written — a DOI in the case of an article, an ISBN in the case of a book, and a report number or an ISSN in the case of a technical report —, and are published by a publisher, except technical reports that are produced by an organization.

5.2. External entities related to a document

Some of these explicit document data indirectly provide two types of information. The first one corresponds to the semantics explained above — for example the authorship of a document —, while the other points to an external and more complex concept. This concept is related at some level with the document in question, but it is clearly different from it. In our model we show as an example three entities — *author, publication venue* and *research institution* — that have this duplicity. In fact, a document typically has the name of the authors explicitly written, which provides a direct information about the authorship of the document, but each of these names, however, also serves as a link to the author himself/herself, and then to a lot more information about him/her such as his/her current involvement in a research institution, which can be different from the affiliation of the author at the time of writing the document. In fact, over time, an author may have been affiliated to different research institutions. Therefore, we model separately the author's name — as it appears written in the document — and the author himself/herself, as the latter concept groups much more information about the author than the first one.

Following the same approach, we also distinguish between a publication venue and a research institution and their respective names. Indeed, any author, publication venue or research institution may have had different names over time. For example, an author can have published some documents using his/her maiden name, and some others using his/her married name. Similarly, a publication venue or a research institution may have changed their name to adapt it to changing times in research.

Moreover, as it happens with a document, we can also distinguish different types of publication venues — *conference proceedings* and *journals* — and research institutions — *research group, educational institution*, and *company* — that are common in the investigation field.

Apart from this, there are three entities — *document, author* and *publication venue* — that can have one or more impact metrics associated — like, for example, how many times they have been cited. These metrics allow CS researchers to quantify the relevance of a document, author or publication venue in the scientific community, or in the context of a physical/digital local collection. This division allows to distinguish between the impact metrics that are calculated with all the public information that is available — like the number of citations, who cites what, who is cited and by whom… —, and those that only take into consideration a subset of this information. As an example, an author can have a high impact metric because he/she has published a lot of relevant documents, but if a CS researcher only considers his/her local collection, the same author can have a lower impact metric, as only a few of these documents —those that the CS research has considered relevant to be stored in the local collection— are used to figure out the impact metric value.

Finally, the document's sub-model illustrates that publication venues specialize in one or more topics, and then usually publish documents related to it; research institutions and authors focus their research in some specific topics; and documents deal with one or more topics. Additionally, some of these authors, publication venues, documents and research institutions not only address some topics, but they can also be a reference in some of them.

Regarding the topics, each of them can in turn be related to other topics — either more abstract, more concrete or complementary— and has a specific terminology associated that is used in the research context. As an example, Wilson (1999b) states that the *"[…] information behavior may be seen […] as a series of nested fields"*, where information

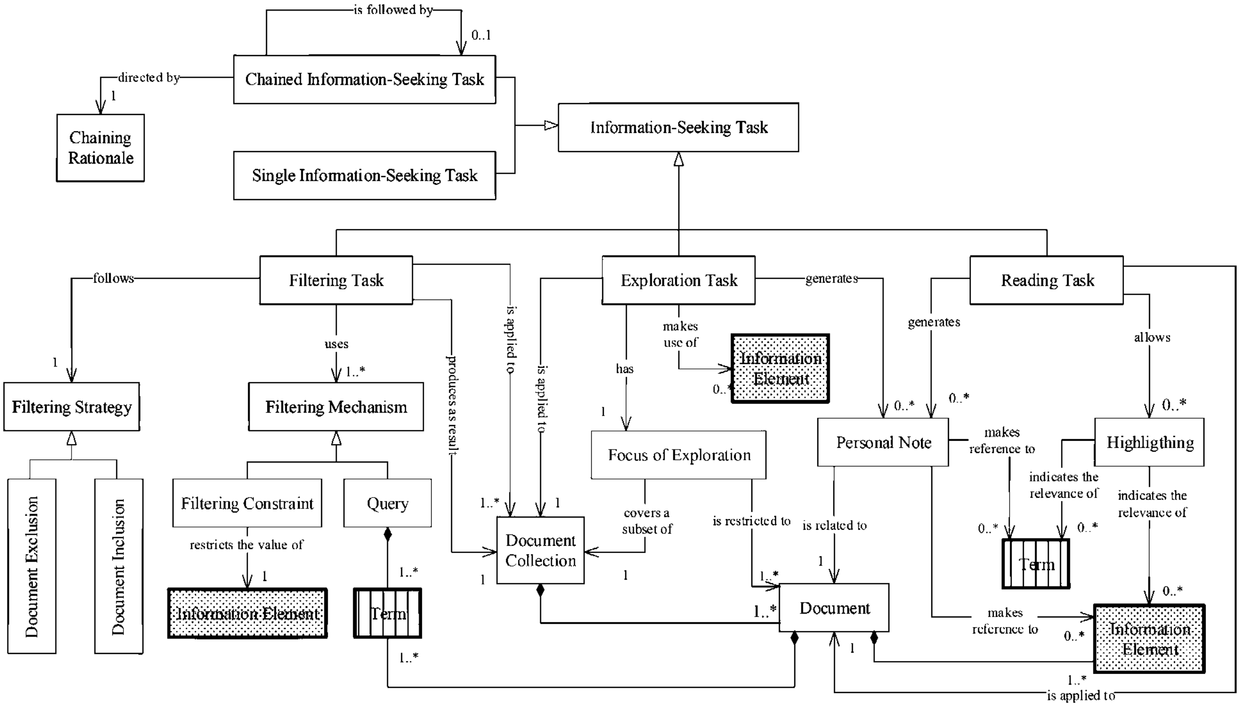

Fig. 4. Information-seeking tasks sub-model. Concepts *Information Element* and *Term* are duplicated in the sub-model for the sake of clarity. They are represented by a specific pattern (outside the UML standard) to highlight that they stand for the same element.

search behavior is a sub-area of the information-seeking behavior, that in turn is a sub-field of the information behavior area.

5.3. Implicit document data

Some of the non-written data that a document can implicitly contain are also represented in this sub-model. To obtain them, the user has to get into the document and read it to some extent – or at least to skim it – in order to process some of its explicitly written information. As an example, after reading a few words of a document, like its title, the reader can identify in which language it has been written. The number of pages of a document is also an implicit datum as its calculation is based on the range of pages, which is an explicit document datum. Besides, by having a quick look at the document, the CS researcher can recognize its structural layout, as scientific documents are usually formed by very well-defined sections like a list of keywords, an abstract, an introduction or a conclusion. In some cases, depending on the information-seeking purpose and the context of the information-seeking activity, a CS researcher may just need to read these sections, as they are – at different levels – summaries of the contribution of the document. In other cases, a CS researcher may just want to have a quick look at the general structure of a document, either through its table of contents or even by quickly overviewing which sections and content elements make up the document and the type of contribution it offers. For example, a CS researcher can be interested in finding documents that include some kind of figures, like graphs or diagrams, whereas another CS researcher can show more interest in a document if it contains some mathematical equations. Occasionally, a CS researcher can also want to skim a document in order to identify its content type. According to interviewees, the contribution of a document may be classified as primary research, when some findings are presented and sometimes validated somehow – for example through theoretical or empirical validation–, but they also as a survey, a review, or a case study, among others.

6. Information-seeking tasks sub-model

After interviewing the participants of our study, we have identified that the actions they perform to seek information can be classified under three main groups. First of all, a CS researcher can need to carry out a filtering task in order to obtain, among a document collection, the sub-set of documents that match to a greater or a lesser extent some criteria. In some cases, a CS researcher may just want to explore a document collection in order to obtain a general idea of its main contents and/or structure. Besides this, in order to achieve a specific information-seeking purpose, a CS researcher may need, at some point, to read, to a greater or a lesser extent, one or more documents.

These information-seeking tasks – whose complete sub-model is provided as Fig. 4 – cannot be seen as stages of a static or predefined pipeline, even if they are complementary. In fact, a CS researcher may start exploring a document collection, then perform some kind of filtering and finally read some of the resulting documents, but it is not the only possibility. A CS researcher may directly select or discard documents from a collection only by reading them — and then without exploring nor filtering the document collection —, or he/she may just want to obtain the list of documents written by a specific author — which can be achieved just by filtering the initial document collection. Besides, due to the variability of needs that a CS researcher may have while seeking information, and to the complexity of the information-seeking process itself, the user has to deal with a great amount of uncertainty (Wilson, 1999a). This implies that the process has to be modeled in the most flexible way possible, avoiding simple and unrealistic one-size-fits-all approaches. To achieve this, information-seeking tasks cannot be considered atomic — that is, a task can start only if the previous one has already finished — as they can interleave. For example, while reading a document or exploring a document collection, a researcher may need to perform a search — which is a type of filtering task — to clarify some aspects he/she does not understand or to expand some information, but, on the contrary, after performing a filtering task, he/she may want to explore the results in order to decide if a further filtering task is needed or not.

In short, a CS researcher may need to perform only a single information-seeking task to achieve his/her information-seeking purpose, but usually he/she requires carrying out a sequence of chained information-seeking tasks in order to iteratively refine the results of an information-seeking activity. This means that the CS researcher performs a set of information-seeking tasks that may be inter-dependent, as the output of each of them — that is the set of documents resulting from a filtering, exploration or reading task — can be used as input in the next task, or at least its results can somehow lead the performance of the next task, for example through the identification of terms belonging to the specific terminology of a topic. As an example, a CS researcher may need to explore a set of documents returned after performing a filtering task, but he/she can also decide to perform a new filtering task after identifying new relevant information during a reading task, or he/she may simply prefer to fully read them. The decision of performing a new information-seeking task relies uniquely on the CS researcher, who has to determine if his/her information-seeking purpose has been achieved. This decision is represented in our model as the *chaining rationale* that guides the tasks chaining. In case of iteration, this rationale also contains the information that is going to be used to link the current and the future information-seeking tasks, and how this information is going to be used. As an example, if a CS researcher wants to perform a systematic literature review, he/she will probably carry out both forward and backward snowballing activities in order to find and process all the documents that are cited by and/or that cite documents considered relevant in the topic being studied. In this case, then, the chaining rationale consists in making use of the bibliographic references — for forward snowballing — and the publication details — for backward snowballing — of relevant documents, until no new relevant documents are found. These features map with the model proposed by Byström and Järvelin (1995), as they assume that the complexity of satisfying an information need is not always the same, and then an iterative and incremental approach is sometimes needed. In their model, the decision of performing another information-seeking task is also defined by the user, who has to evaluate if his/her information has been satisfied.

6.1. Filtering task

Due to the huge amount of information — both scientific and generic — that is currently available in the Web, when a CS researcher needs to find some specific information to achieve an information-seeking purpose, he/she often has to reduce the number of documents contained in the collection to be explored in order to keep only those that are potentially relevant for his/her current purpose.

A CS researcher can employ one or more filtering mechanisms in order to filter a document collection, but he/she typically has at his/her disposal only two filtering strategies. Both lead to the same result, but using a different approach. In the first case — *document inclusion* —, the CS researcher identifies which documents are potentially relevant to achieve his/her information-seeking purpose, and then have to be kept. The second strategy — *document exclusion* — consists in discarding the documents that are not relevant nor useful to achieve the current CS researcher's information-seeking purpose. In the first case, only the documents meeting the criteria defined in the filtering mechanism are considered potentially relevant, while in the exclusive case, documents meeting these criteria are those that get discarded. For example, a CS researcher can decide to discard all the documents that are not written in English, while another CS researcher can decide to keep only the documents containing graphs, diagrams and/or tables as he/she looks for illustrated quantitative results. Even if these two strategies seem to be equivalent and can lead to the same results, they are different in terms of how the CS researcher faces the filtering task — see more details in Section 7.

A filtering mechanism can be implemented as a query or as a constraint. A query is typically used to retrieve documents from an information source in the classical best-match approach, where the use of a search engine is required in order to obtain only the documents that contain the query terms — maybe not all of them, but most of them. On the other hand, a CS researcher may also prefer to filter a document collection by defining one or more filtering constraints. Each of these constraints indicates which pair of information element and specific value — or range of values — need to have the documents to be affected by the filtering — either to include or to exclude them. For example, a CS researcher may want to keep only the documents that have been published in a specific journal, or, on the contrary, he/she may want to exclude those documents published before 2010. Obviously, a CS researcher may also want to combine both mechanisms — that is defining both a query and one or more filtering constraints — to perform a more narrowed and detailed filtering task. For example, a CS researcher may want to obtain only the documents that match a given query and that have been published after a given date.

6.2. Exploration task

As opposed to the filtering task, an exploration task does not aim at selecting one or more specific documents or at reducing a document collection, but just at exploring it. When exploring a set of documents, a CS researcher does not have any particular objective beyond, for example, getting a general idea of which type are these documents, which are the main topics they address, or who are the most recurrent authors, among others. In order to do this, a CS researcher can manually explore one-by-one all the documents, but he/she can also perform some kind of ordering in order to create clusters according to one or more information elements. Then, the exploration is not directed by straightforward and well-defined criteria — as it occurs when using a query and/or a set of filtering constraints —, but instead by the dynamic decisions and intuition of the CS researcher while browsing the document collection. Therefore, the CS researcher is the one who decides at every moment how to proceed with the exploration and which information he/she needs to collect and use. In fact, many of the interviewees mentioned that, during their exploration tasks, they usually take some personal notes about everything they consider relevant or useful to achieve their information-seeking purpose. This relevant information can consist of any of the information elements contained in the document — like the name of its author(s), the name of the venue where it has been published, an equation, or a table —, but also of a set of terms contained in the document that the CS researcher considers relevant — for example because they are part of the specific terminology of the topic addressed in the document. In both cases, the CS researcher usually wants to clearly identify the most relevant content, either for having a summary of it — for example for speeding up further exploration or readings of the document —, or for using it as part of the chaining rationale in following chained information-seeking tasks. Besides, a personal note can be related to different documents as a CS researcher may want to relate different documents for some reason.

Given that in the exploration task no filtering is performed, it may be difficult — or even impossible — for a CS researcher to explore big document collections, and that's why he/she typically has to focus the exploration on a reduced subset of the document collection at every moment. Then, for example, if a collection is formed by a hundred documents, the CS researcher may decide to explore them by tens.

6.3. Reading task

At some point, a CS researcher may want to open and read a document — either fully or partially. While reading a document, a CS researcher can identify some relevant information and may want to highlight it — for example by marking it with some kind of symbol, by underlining it, or by coloring it with a physical or digital highlighter — or to take some personal notes about it. In this case, additionally, a CS researcher can take some personal notes about the document as a whole, but also about aspects that are not explicitly included in the document but that may have raised during or after reading it, like new promising ideas that have not been considered previously, or new approaches for solving a problem.

7. CS researcher sub-model

Undoubtedly, the concept that introduces most complexity in the description of the information-seeking process is the CS researcher himself/herself, due to the difficulty associated with trying to model human characteristics and behaviors — see full CS researcher sub-model in Fig. 5. First of all, each individual has his/her own demographic profile including personal information like his/her gender, nationality, age or hand's laterality — that is if he/she is right-handed, left-handed or ambidextrous. Additionally, each CS researcher is unique in terms of professional background, personal preferences, and personal characteristics, both physical and psychological. Within the latter category, we can distinguish between cognitive characteristics —there are different models of cognitive or learning styles, for example in terms of field-dependence — and personality traits — for example how much self-confidence the CS researcher has while performing an information-seeking activity.

How and why all these aspects can affect the performance of an information-seeking activity would require a more specific and detailed study of CS researchers when they seek for information, taken from a psychological point of view. That is why, even if we have some intuition about these cause-effect relationships — especially related to the experience and confidence of CS researcher —, we only describe the concepts mentioned by the interviewees and let a more in-depth analysis for future work.

Unlike relevant knowledge and personal characteristics that are inherent to CS researcher and then independent from other concepts, there are two types of personal preferences. A CS researcher can have some predefined personal preferences according to his/her tastes, capacities, abilities, and so on. Additionally, while performing an information-seeking activity, a CS researcher can also have some current preferences that may be different as they are influenced by the specific information-seeking purpose he/she wants to achieve and the information-seeking tasks he/she performs, and/or by the context in which the information-seeking activity takes place. Due to this substantial difference, that may also be related to the natural evolution of the user as person, researcher, and information seeker, it is essential to capture all the actions performed by the CS researcher and their consequences in order to properly model him/her. Then, it is required to keep the

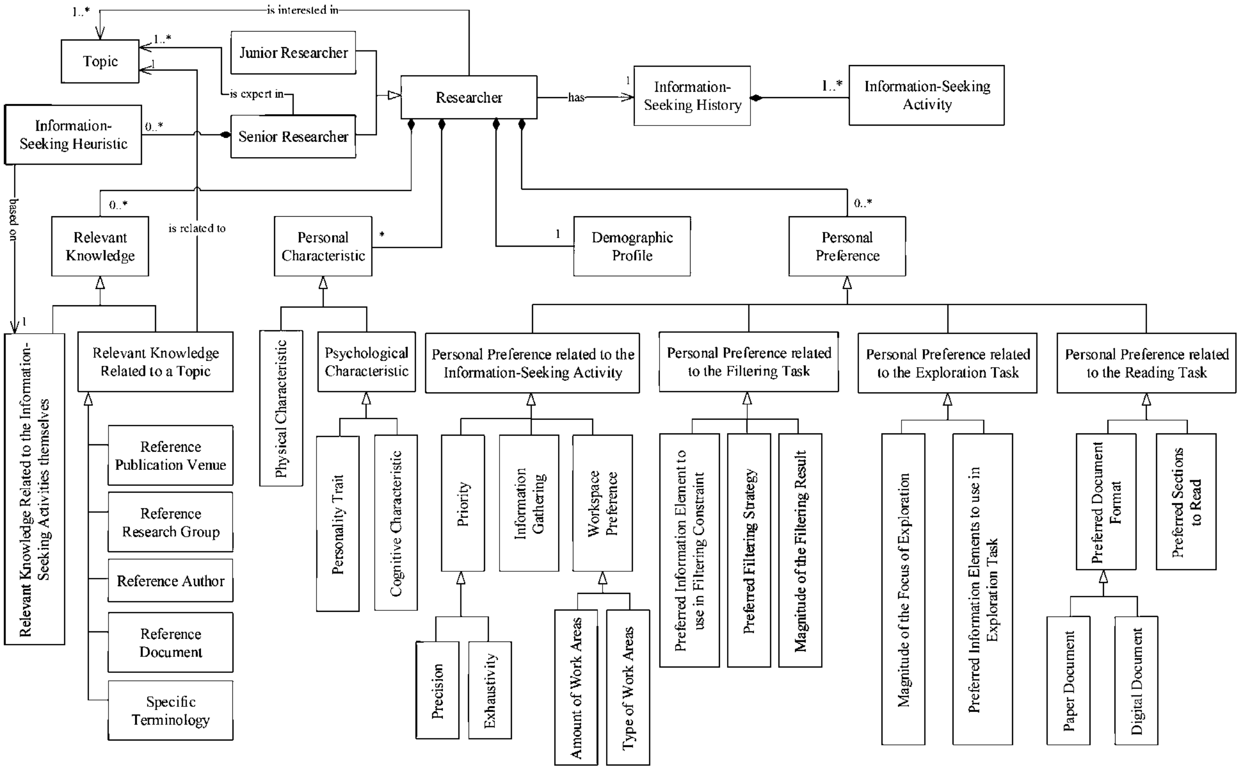

Fig. 5. Computer science researcher sub-model.

track of the information-seeking activities performed by the CS researcher, and of all the circumstances that surround them, like for example, which documents he/she has already consulted, stored or discarded, which are the terms most frequently used to build a query or which are the sources of information he/she consults more recurrently.

At a psychological level, the analysis of the qualitative studies reflects that intuition and self-confidence are two of the personal aspects that define the CS researcher's behavior when performing an information-seeking activity. A CS researcher gradually gains self-confidence and resorts more to his/her intuition —almost unintentionally — as he/she undergoes information-seeking activities once and again. Even if this is a rough simplification of a much more complex construct that goes beyond the scope of this article, we define a CS junior researcher as the one who lacks — or has very little — previous experience in seeking and finding information in a research context. This often translates to performing more information-seeking tasks than those actually needed in order to achieve an information-seeking purpose, and almost systematically to doubt the goodness and relevance of the obtained results. Conversely, a CS senior researcher is supposed to have a lot more knowledge related to the information-seeking activities themselves as he/she has performed many of them throughout his/her career. In fact, most of the interviewed CS senior researchers were able to more clearly explain which steps they carry out when seeking information, as they follow a more standardized process based on their experience.

In practice, this procedural knowledge translates in the acquisition and use of some heuristics — even if the CS researcher is not even conscious of it — that facilitate and speed up the mentioned information-seeking activities, and that improve the quality of the results. Initially, many of the interviewed CS senior researchers instinctively thought they do not use heuristics. Then, even after thinking about it for a while and realizing that they actually have and use some of them, they found it difficult to explain them or even to make them explicit. This reflects that these heuristics are very internalized by the CS researcher, that uses them systematically and unconsciously. As an example, one of the interviewed CS researchers explained that she usually considers only the first 15–20 documents of a filtering task result, ordered according to their relevance or to the closeness to the query, as, according to her experience, beyond this number it is highly unlikely to find documents that are actually relevant for her current purpose.

Additionally, even if a CS researcher — whether senior or junior — is typically interested in one or more specific topics, senior CS researchers usually are also experts in at least one of them, as during their research activities they have obtained a great amount of knowledge about them.

7.1. Relevant knowledge

One of the CS researcher's aspects that most influences the information-seeking activity is the amount of relevant knowledge he/she has. As explained just above, part of this knowledge is related to the information-seeking activities themselves, while the other part is related to a topic, as a CS researcher can already have some knowledge about the investigated topic, like who are the reference authors, which are the reference research groups that investigate on it, which are the reference documents and/or publication venues dealing with it, or which is its specific terminology.

This knowledge can be used, for example, to define the specific values of the filtering constraints used in the filtering tasks or to decide on which documents should the CS researcher focus his/her attention. Indeed, there are many CS researchers that are considered experts, either because they are CS senior researchers and then are proficient undertaking information-seeking activities, or because they have a lot of knowledge about certain topics — like a Ph.D. supervisor, the author of a prominent research document, or a conference keynote speaker. These experts can help to gain access to documents that are relevant for the current purpose by sharing their relevant knowledge related to a topic or by providing some useful heuristics that can facilitate the performance of the information-seeking tasks. In this regard, it is important to distinguish between expertise and relevance, as a CS researcher can be an expert in a topic, but yet it may not be considered a reference author in it by the scientific community.

7.2. Personal preferences

As mentioned before, each CS researcher is different and has different personal preferences that fit with his/her particular and individual point of view when it comes to perform an information-seeking activity. In general, a CS researcher has a set of predefined personal preferences — depending on his/her profile, tastes, experiences, and so on. However, at a given moment, these preferences can change in order to adapt to his/her current needs and to the restrictions that may exist, mainly because of the context in which the information-seeking activity is carried out — see Section 8. The personal preferences that the participants of our qualitative studies have directly or indirectly mentioned are presented below. They are probably a small sub-set of all the possible personal preferences that may exist but, as mentioned in the discussion section, our model is flexible enough to accept new concepts and sub-concepts, like personal preferences, without affecting at all the rest of the model.

7.2.1. Preferences related to the information-seeking activity in general

Beyond the needs that are implicitly associated to each information-seeking purpose, a CS researcher may have some personal preferences related to the information-seeking activity in general. To be more specific, we have identified that some CS researchers have as a priority — in general terms, and regardless the context or the information-seeking purpose — to obtain very precise results, while others may prefer to obtain exhaustive results. In the first case, the CS researcher prefers to obtain all the documents that can be useful at some level to achieve his/her information-seeking purpose, even if this may imply obtaining also some documents — many or few — that are not actually relevant for his/her current purpose. In the second case, a CS researcher may prefer to obtain a small subset of results that are indisputably relevant for the current purpose, even if this may imply missing some documents that could also be somewhat relevant. According to the performed interviews, CS senior researchers usually prefer to be precise, while CS junior researchers use to feel more confident when they are exhaustive.

Moreover, each CS researcher prefers to gather different types of information to follow up with the information-seeking activity, like for example:

- the terms that the CS researcher plans to use to refine the query in the following chained information-seeking task,
- the information elements that, according to him/her, he/she should use to successfully explore a document collection,
- or the terms and information elements that actually allow him/her to achieve his/her information-seeking purpose.

Finally, a CS researcher may have some preferences about where to perform the information-seeking tasks. More concretely, he/she may be used to work with a given amount of different work areas — for example a computer with two screens — and/or to perform some of the information-seeking tasks in a specific support — for example reading a conference article in a tablet.

7.2.2. Filtering preferences

As mentioned before, the primary objective of a filtering task is to reduce the size of a document collection, and this can be done by selecting those documents that match the criteria associated to the current purpose, and/or by discarding those ones that do not match them. Which of the two strategies is a priori adopted depends on the CS researcher preferences. According to the results of our study, CS researchers usually follow one or another strategy according to their expertise, and above all to his/her degree of self-confidence. In fact, for a CS junior researcher it seems to be a lot more difficult to ensure that a document is not relevant for a given purpose than for a CS senior researcher, as the former has the tendency to feel that everything is important and then fears discarding something that might be relevant. CS senior researchers, on the contrary, have a more critical and confident judgment when determining if a document is relevant or not.

On the other side, we have mentioned that during the filtering task, the selection or discarding of a document relies on the value of one or more of its information elements. Again in this case, each CS researcher can prefer to use some information elements in order to define the filtering constraints. As an example, one of the interviewees affirmed that she usually can determine if a document is potentially relevant for her current purpose just by knowing its title and authors —

then, many times she only considers them to select or discard documents —, while many other participants expressed that they always take into consideration, at least, the title, the abstract, the introduction and/or the conclusion.

The last personal preference related to the filtering task has to do with the chaining rationale explained in Section 6, as it is related to the size of the document collection being manipulated. A CS researcher may want to perform as many chained filtering tasks as needed to reduce the number of resulting documents enough to being able to explore it more in detail. To what extent the document collection has to be reduced —that is, which is the *magnitude of the filtering result* — in order to stop filtering a document collection depends on the information-seeking purpose to be achieved, but mainly on the CS researcher preferences, as he/she can prefer to explore a rather small amount of documents, or to process a higher set of documents. As an example, a given CS researcher might feel comfortable exploring a set of a hundred documents returned after filtering a collection, while another one might feel overwhelmed by such amount of documents and may prefer to further refine the filtering result so that he/she gets no more than twenty documents.

7.2.3. Exploration preferences

As mentioned before, when exploring a document collection, even if a lot of documents are displayed at a time, a CS researcher typically only draws his/her attention to a small subset of them at every moment in order to avoid being overwhelmed by too much information. Thus, how much information or how many documents should be displayed simultaneously — that is, which is the *magnitude of the focus of exploration* — also varies from one CS researcher to another.

On the other hand, while exploring a document collection, a CS researcher may prefer to use some information elements whether to order the documents, or to determine which of them are potentially relevant for his/her information-seeking purpose.

7.2.4. Reading preferences

Typically, a CS researcher reads the whole document to determine if it is actually useful for his/her information-seeking purpose but, occasionally, he/she may only need to read a small part of it. In these cases, the CS researcher can prefer to give priority to some sections as he/she may consider they contain the information he/she needs to achieve his/her purpose.

Finally, a CS researcher may prefer to read a document in a paper format or in a digital format. In fact, many of the interviewees expressed that they perform the whole information-seeking activities in a digital context but, when they have to read a document — especially an article —, they prefer to work on a physical document as its manipulation is more natural and they can manually annotate and scrawl the content without any restriction.

8. Context sub-model

Finally, there are some external aspects that have to be considered when modeling the information-seeking process. These aspects define the context in which an information-seeking activity is performed. The complete sub-model of the context is shown in Fig. 6.

8.1. Interaction detail

In order to perform an information-seeking task, a CS researcher has to interact somehow with the information system and with the documents. How this interaction is performed is one of the components of the context. To represent this interaction, it is required to know, first of all, which type of device is going to be used by the CS researcher during the information-seeking activity, both for inputting and getting information from the system.

On the other hand, it has to be defined which means of access is the CS researcher going to use to filter, explore and/or read the documents. The most basic and elementary means of access is the manual access, where the CS researcher filters, explores and reads a set of documents without using any technological tool. In this case, the CS researcher directly interacts with the documents, usually belonging to a local document collection, in order to achieve an information-seeking purpose. This means that he/she navigates through them by manually accessing the folders that, according to their name — for example —, can contain documents relevant for the current purpose. This type of access is the one used when the information-seeking activity makes use of physical collections of documents, but also when software is only used to visualize the documents — like the file explorer of an operating system does, for example — and the information-seeking tasks are performed manually by the CS researcher. As an example, in order to filter a document collection, instead of explicitly indicating a query or a filtering constraint in a software-based tool, the CS researcher is the one deciding if a document or a set of documents can be considered potentially relevant for the current purpose. This, for example, is quite common when the CS researcher looks for a book in a library. In practice, manual access can only be performed if the document collection is rather small, as it requires from the CS researcher to spend a high amount of time to manually and individually look over all or most the documents of a collection.

In fact, research document collections usually contain a huge amount of items, and then it becomes essential to use a software-mediated tool to perform the information-seeking activities in a reasonable amount of time. A CS researcher can use a search engine — for example — to automatically filter a document collection based on a set of filtering criteria, or to find where a specific term has been used in a document in order to narrow the reading to this concrete part of the document. But he/she can also use an information visualization system, usually integrated in an information system, to

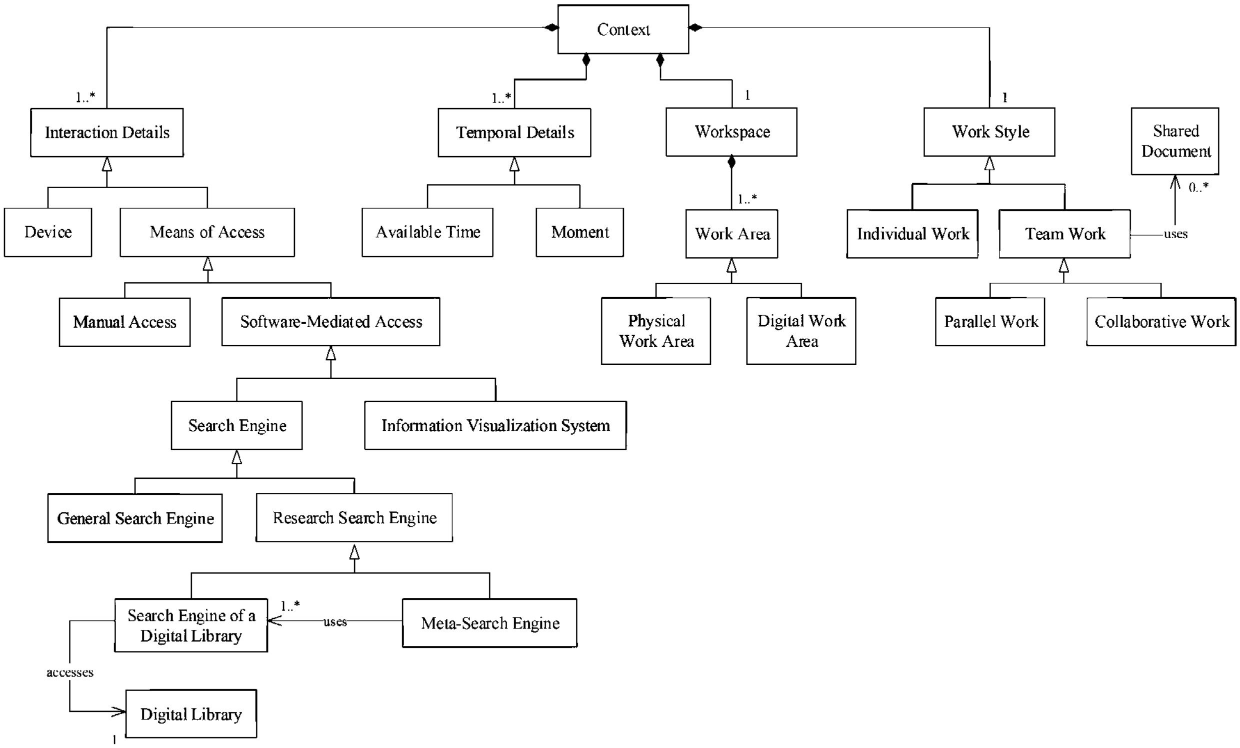

Fig. 6. Context sub-model.

facilitate observing the results of a filtering task, exploring a set of documents by allowing to visually classify them using a set of visual characteristics – like the color, the size, or the shape –, or annotating a given document while reading it.

In the specific case of the search engines – that are indisputably one of the digital tools that are most used in research –, there are two kinds that can be used depending on the aim of the search, but also on the type of information expected to be found, on the previous knowledge of the CS researcher about the topic, and/or on his/her expertise in carrying out information-seeking activities. On the one hand, general search engines like Google, Yahoo! or Microsoft Bing are typically used to obtain a wide variety of information, even if it does not stem from scientific research sources. This kind of search engines are often used by CS researchers in rather initial exploratory phases of an information-seeking activity, mainly because they intend to expand the scope of the search as much as possible; because they have little or no knowledge about a topic; because they do not know exactly which document collections could contain documents relevant for the current purpose; or because they still do not know the specific terminology used in the topic that would allow them to define good search criteria.

On the other hand, there are many search engines that are intended for research purposes. Some of them are strictly associated to a single digital library, while others, called meta-search engines, provide a conceptually higher layer that allows a CS researcher to transparently perform the same search over more than one digital library, but using the same centralized interface. This is possible because, internally, a meta-search engine performs the same search over many specific search engines and integrates all the obtained results. These results are then presented in a homogeneous way, even if each of them points to its corresponding digital library, which facilitates and simplifies its use. Research search engines are usually employed by CS researchers that already have some relevant knowledge related to the topic – like its specific terminology – and then can perform a more directed search.

8.2. Temporal detail

The second type of attributes of the context are those related with time. This implies knowing the moment when the information-seeking activity is carried out, and how much time has the CS researcher available to perform the activity. In the first case, it is important to know when the information-seeking activity takes place because there are some aspects, like the CS researcher's level of fatigue, that can influence his/her performance, concentration, motivation, and so on. As an example, a CS researcher is probably more tired late in the afternoon than early in the morning, and then in the first case he/she could be more distracted and perform worst. On the other hand, the amount of time that the CS researcher can spend in seeking information can also affect the performance of the information-seeking activities as, for example, it can narrow the amount of results the CS researcher is able to explore, or the amount of chained filtering tasks he/she can carry out to refine the result as much as possible according to his/her information-seeking purpose.

8.3. Workspace

Alongside with the moment when the information-seeking activity is carried out, it is also important to take into account the workspace where it takes place. A CS researcher can work in many different workspaces — for example at home, in his/her workstation, in a laptop while flying to a conference... Additionally, a workspace can in turn be formed by many work areas that can be categorized as digital work areas — for example a computer or a tablet— or physical work areas — for example a sheet of paper. Usually, a CS researcher uses each of these areas for a different activity, like searching, exploring, reading or writing. As an example, a CS researcher can use two screens while seeking information, using one of them as a work area to perform the filtering tasks, and the other one to take notes about what he/she is reading. Alternatively, he/she can take some notes in a paper notebook and, at the same time, make use of a digital word processor to copy and paste relevant statements or references. Analogously, a CS researcher can filter a document collection using a search engine, but he/she can also manually look for a specific article in a printed journal issue, for example.

8.4. Work style

Finally, a CS researcher may seek information individually, but he/she also may be part of a research group, and then it is highly probable that he/she has to coordinate with other CS researchers to carry out the information-seeking activities. In this case, the team can work collaboratively so that each member performs a separate part of an information-seeking activity and then they integrate their respective results. On the other hand, the team can work in parallel so that each member seeks the same information — but probably by different means, using a different sequence of information-seeking tasks, and using, for example, different mechanisms to filter the collection, or different information elements to cluster the documents — in order to dynamically share and complement their results.

In both cases, the team shares the same information-seeking purpose, but in the first case each member is in charge of fulfilling a different part of it, while in the second case a more holistic approach is undertaken and everybody tries to achieve the whole information-seeking purpose. Then, for example, if the information-seeking purpose is to elaborate a state-of-the-art of a topic, each member of the research team can take care of looking for relevant documents published within a specific range of years so that, altogether, they cover they whole desired period. Alternatively, all of them can perform the information-seeking activity considering the whole period of time. In terms of information retrieval, both approaches differ in terms of precision and recall, that are the most common measures used to assess the effectiveness in this field (Manning, Raghavan, & Schütze, 2008). Collaborative work can potentially provide a higher recall as the number of documents that have to be processed by each CS researcher is smaller and then it is more difficult to let a relevant document unidentified, but the precision can be lower as determining if a document is relevant or not depends only on the criterion of one CS researcher. On the contrary, the results obtained by teams working in parallel can potentially be more precise as the findings can be contrasted: if all or most of the CS researchers have found the same document and have considered it relevant, then it is almost sure that the document is actually relevant. In this case, however, it is more probable for some relevant documents to be unidentified as the document collection is bigger, and then the CS researcher potentially has to perform more information-seeking tasks and spend more time in each of them.

In both cases, CS researchers that are part of a team usually establish one or more shared documents — for example a digital word processor's file stored in the cloud — as a meeting point where they write their comments, results, summaries and/or interesting ideas, so that the rest of the team members can be aware of them.

9. Practical applications of the models

We believe that the model can be of great utility in different contexts. First of all, it can be used as a framework for characterizing existing information systems addressing information-seeking for research purposes, by providing a set of concrete, well-defined, non-ambiguous and specific terminology related to the process. It can be also used for comparison of different competing systems for information-seeking activities in research. For example, the model can be used to describe which information-seeking tasks Google Scholar allows to perform, and which are its main characteristics. Fig. 7 illustrates this visual description by green coloring the concepts of the information-seeking tasks sub-model that intervene at some point in an information-seeking activity performed in Google Scholar.

As we can see, Google Scholar allows filtering, exploring and reading documents. In the case of the filtering task, for example, we can also see that Google Scholar allows using both filtering mechanisms. On the one hand, the user can pose a query formed by a set of terms in order to obtain as filtering result the list of documents that match with these terms. This task is applied to an initial document collection containing all the documents indexed by Google Scholar, and produces a subset of them. Additionally, the user can define a set of filtering constraints like the document collection where the filtering will be applied upon — all indexed documents, case law or personal library —, the year or range of years when documents must have been published to be considered, and if certain types of documents have to be considered — patents and citations. All these mechanisms follow an inclusive strategy, as they define which characteristics needs to have a document to

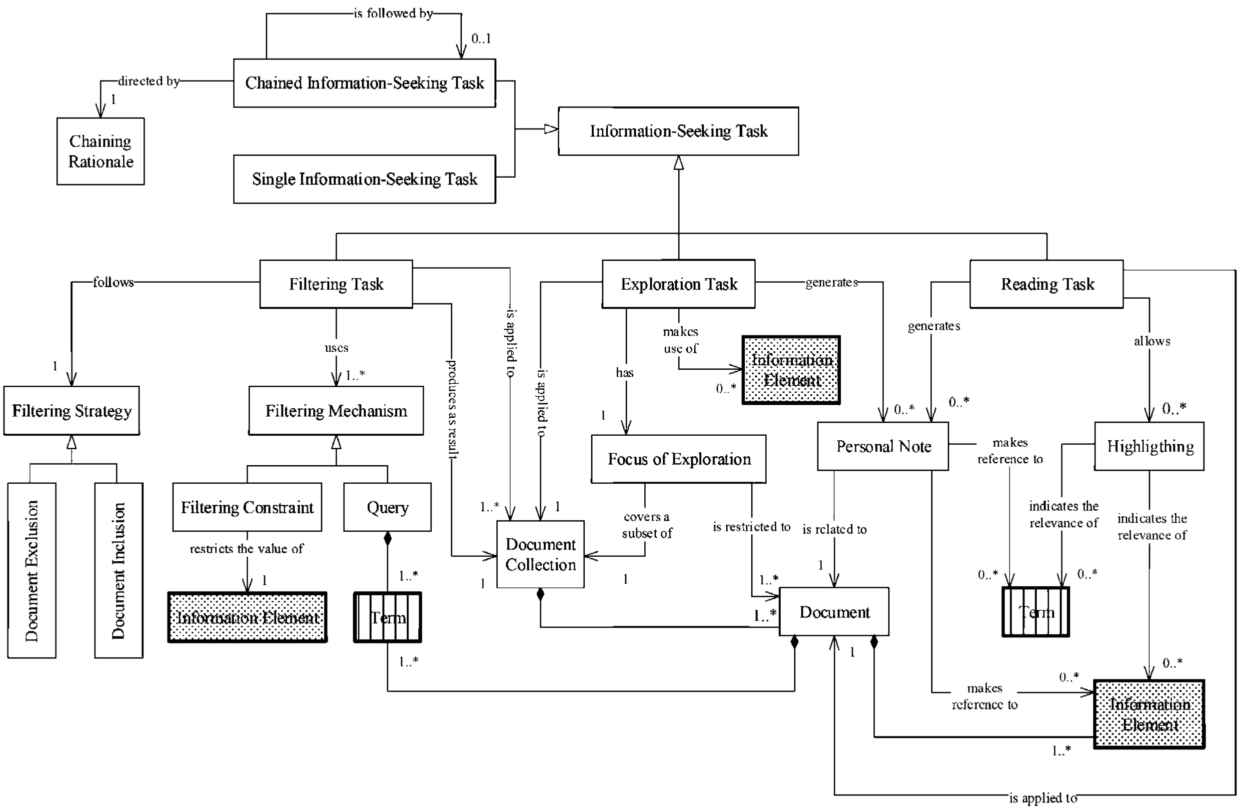

Fig. 7. Description of the information-seeking tasks supported by Google Scholar using the conceptual model.

be included in the resulting document collection. Exploration, in turn, is performed over a specific document collection that is always the one resulting from a previous filtering task. To be more specific, Google Scholar allows to perform the exploration in groups of ten or twenty documents, so that the user focus the exploration only on a small subset of documents. This implies that Google Scholar allows performing chained filtering tasks aiming at refining the resulting collection of documents iteratively and incrementally, interleaving diverse tasks if needed. Of course, Google Scholar also allows performing a single task without any further refinement. Finally, Google Scholar also allows reading the documents belonging to the filtering result collection and that at that moment are included in the focus of exploration. Nonetheless, Google Scholar does not allow to highlight relevant parts of a document, nor adding personal notes to them.

As an extension of this use, the models can be used to compare different existing information systems. For that purpose, all the concepts intervening in the activity have to be instantiated in order to be able to perform the comparison at a lower level. Then, for example, in the case of Google Scholar, the *filtering constraint* concept should be instantiated by the concrete values *source of documents, publication date*, and *content type*, and this would allow to determine if it has the same, more, or less *filtering constraints* than other information systems.

Beyond these possible uses, having this model may be very useful when designing an information system if a user-centered design approach wants to be followed (Preece et al., 2015). This approach is based on the observation and elucidation of the current context of use — formed by the user, the tasks and the environment — of the system being designed. The conceptual model allows to formally represent the analysis of the information provided by real potential future users during the qualitative studies. Besides, one of the most intrinsic characteristics of the user-centered design is that the process is iterative and incremental, as solving complex tasks, like seeking information in the research field, requires increasing the knowledge about the problem along several iterations, and incrementally designing a system addressing the identified problems and evaluating at every iteration if the solution already solves the problem properly, or if there are aspects that have been poorly addressed in the solution or that have not been covered in the analysis of the context of use. Our model, being flexible, adaptable and expandable, perfectly adapts to this philosophy and then is a perfect tool for specifying the context of use of an information system used by CS researchers when seeking information related to their investigation.

Finally, the models can be used as a standard terminology allowing to comprehensively describe specific information-seeking activities without forgetting any important aspect. This is essential for describing and comparing existing information systems, but also to develop useful and complete support material to train novices in seeking information in a research domain, for example.

10. Discussion

As explained in the literature review section, Bates (1989) states that it is needed to model the user, the tasks, and the context in order to properly model the information-seeking process. Our work matches these findings, as we consider that there are five main aspects that influence the performance of an information-seeking activity by a CS researcher, as reflected in the main model described in Section 4 and illustrated in Fig. 2. First of all, the purpose to be achieved by a user when seeking information has to be considered and included in the model. Apart from it, we have also included in the model two of the main aspects that intervene in the activity: the documents used by the CS researcher as source of information, and the CS researcher himself/herself. Both aspects have been respectively developed in Sections 5 and 7, and represented through the sub-models illustrated in Figs. 3 and 5. The two remaining aspects that intervene in an information-seeking activity, and then have to be modeled, revolve around the user: the tasks he/she needs to perform to achieve his/her purpose, and the context in which he/she the information-seeking activity takes place. Both aspects have also been explained and modeled separately in Sections 6 and 8 – and illustrated in Figs. 4 and 6 –, respectively, in order to provide as much details as possible.

This approach has also been supported by other authors, like Byström and Järvelin (1995), who argue that modeling the information-seeking mainly relies on the user, or Domik and Gutkauf (1994) that state that there are four aspects that have to be considered to model the information-seeking process. In fact, most of these aspects have also been included, to a greater or lesser extent, in our model. First of all, the "*Data model*" proposed by and Domik & Gutkauf is comparable to our "*Document*" sub-model. Then, they also state the user has to carefully be modeled, as we also do –through the "*User model*" in their case, and the "*CS Researcher*" sub-model in our case. Finally, the "*Resource model*" of Domik & Gutkauf directly maps with part of our "*Context*" sub-model, as it deals with the hardware and software used by the user to carry out the information visualization process. Finally, even if their "*Problem domain/task model*" is not explicitly included in our model because it is specifically related with the visualization of the information, it poses as principle that the visualization of the information obtained through information-seeking activities clearly depends on the domain in which they are carried out. In this work, as we will explain below, we agree with this assertion and we consider that the domain in which the information is being is sought has to be taken into account when modeling the process.

In relation to this aspect, even if some authors have proposed general models that are supposed to be able to describe any information-seeking activity, regardless the domain in which it is carried out (Bates, 1989; Cheuk & Dervin, 1999; Godbold, 2006; Marchionini, 1997; Wilson, 1997), many others have stated that the specific domain in which the activity is carried out is an aspect that has also to be considered in the model (Cheuk & Dervin, 1999; Ellis, 1989a, 1993; Ellis & Haugan, 1997; Ellis et al., 1993; Freund et al., 2006; Kuhlthau, 1991; Meho & Tibbo, 2003; O'Brien & Buckley, 2005; Buckley et al., 2006; Makri et al., 2008; Benardou et al., 2010). This has also been our premise, and then our model is intended to conceptually describe the information-seeking activities performed by CS researchers. This does not mean that the model is not valid or useful to describe the activities performed in other domains, but only that its validity in these cases has not been sought nor verified. We plan to do this validation in future works.

We have also explained that, in the literature, the information-seeking process has been modeled several times as a sequence of stages (Ellis, 1989a, 1993; Ellis & Haugan, 1997; Ellis et al., 1993; Makri et al., 2008; Meho & Tibbo, 2003). In all the cases, authors have assigned to some level a dependence between these stages as there is some kind of precedence when carrying them out. However, all of them also argue that the process cannot be perceived as a linear process where a set of atomic and sequential tasks are performed to fulfill the information need, but instead it must be described as a non-linear and iterative process with a great complexity that has to be iteratively reduced through a set of tasks that may be carried out in an unordered fashion, that may complement each other, and that may even be performed simultaneously. In our model, we also reflect this complexity and uncertainty by avoiding defining the process as a sequence of stages or tasks, from which some kind of precedence or dependence could be derived. However, it is essential to note that, despite this structural difference in the model, most of the concepts presented by these authors have been also derived from the analysis of our qualitative studies – for example *chaining, browsing, differentiating, extracting, filtering, accessing*, or *examination of results*.

The literature review has also reflected that some of the most relevant models related to information-seeking have focused on the user and on the behavior he/she has during the process (Kuhlthau, 1991; Wilson, 1981; Marchionini & White, 2007; Foster, 2004). Even if our model shares many of the concepts presented in these behavioral models, our approach is totally different. We agree on the importance of taking into account the specificities of the user during the process, but our analysis does not focus on how the user feels, what are his/her thoughts during the process, or how he/she acquires and consolidates knowledge during the process. Our proposal follows a more holistic approach in order to try to clearly and unambiguously illustrate by visual means which are the main concepts that intervene in an information-seeking activity.

With regard to the representation of the models, most of the authors have proposed a textual representation, but there are also some authors that have used a pictorial representation (Leckie et al., 1996; Niedźwiedzka, 2003; Wilson, 1981, 1997). In all three cases, nonetheless, even if some concepts related to the information-seeking process are presented, their relationships are defined in terms of precedence, and therefore they better can be defined as sequence or activity models allowing the reader to navigate through the temporal sequence stages through which a researcher has to pass when seeking information. Our conceptual model, however, has to be timeless as it intends to illustrate the relationships as semantic and functional interactions between concepts – reflected by association relationships –, where dependence between concepts is only reflected by structural relationships —like in the composition and generalization relationships. As already explained, this approach has also been followed by Krikelas (1982) and Benardou et al. (2010), who have used a conceptual model to describe the information-seeking activities. However, in both cases, the information-seeking process is over-simplified and the presented concepts are so abstract that the model is not comprehensive nor self-sufficient, and then the reader may have several questions about the model that remain unasked even after reading the explanations provided by the authors, as the concepts and their relationships are not always explained in very much depth. In our opinion, we can intuitively state that visually representing a model is probably easier and faster than explaining it extensively, especially if the final aim is to allow the reader creating his/her own mental models about a given situation or process. For that reason, our visual model tries to be not only comprehensive, but also absolutely self-explainable and self-contained, so that reading the textual explanation is a complementary task whose only aim is to facilitate and improve the understanding of the model by providing more in-depth details and some examples illustrating the concepts and their relationships.

Finally, as stated before, there are several models that have been derived in various fields of engineering (Pinelli, 1991; Holland & Powell, 1995; Leckie et al., 1996; Ellis & Haugan, 1997; Hertzum & Pejtersen, 2000; Anderson et al., 2001; Kwasitsu, 2003; Freund et al., 2006), or from academic or industrial research scientists (Ellis, 1989b, 1993; Ellis et al. 1993; Foster, 2004; Meho & Tibbo, 2003), but there are no models of the information-seeking process performed by the researchers with a strong technical and technological background, like CS engineers. The only models found in the literature review that are somewhat related to this specific user profile are those proposed by O'Brien and Buckley (2005) and Buckley et al. (2006), as they describe the information-seeking process performed by computer programmers. However, programmers, even if they have a strong background in technical aspects and in information and communication technologies, cannot be considered CS researchers, as they have different objectives and focus on different aspects —in concrete, the type of information-seeking purposes they have to achieve are quite different, and the type of documents and information they seek are also different. This, together with the incredible advancement of the technology in the last years, has motivated us to study how CS researchers — that are supposed to master the use of technology — actually perform their information-seeking processes, and evaluate if there are differences with respect to researchers investigating in different domains. As we state in the conclusion section, in the future we plan to formally evaluate to what extent our model is valid in other research fields.

Then, we claim that our contribution can be very interesting as nobody, to the best of our knowledge, has conceptualized the information-seeking process in the specific domain of CS researchers so comprehensively in order to provide a solid, flexible and complete framework.

11. Limitations

Even if we have tried to thoroughly analyze all the information provided by the qualitative studies participants, it must be taken into account that the results have to be contextualized. First of all, these results are directly linked to the profile of these participants, especially with regard to their research domain, computer science. Despite our efforts to abstract and generalize our conclusions, we acknowledge that our model is domain-dependent.

Likewise, we do not intend to suggest that the model is exhaustive as, in general, it represents the aspects raised by participants in the qualitative studies. The authors have interpreted the gathered information, only including their personal understanding of the concepts being organized in the model so as to provide a coherent result. Thus, there is a limited risk that authors have included their own knowledge and beliefs in the resulting model. But in no case the authors have tried to cover all the types and subtypes of the concepts that may exist in research-related contexts. As an example, in our model we reflect that a research document can be of nine types — journal article, conference paper, master thesis, Ph.D. dissertation, compendium of documents, text book, technical report, webpage or forum post — but obviously there could be more, like conference poster, that could be added to the model by a new generalization relationship without affecting its consistency and validity. Using this same example, the authors have included the concept "*document*" in order to generalize the previous concepts into a unique concept, allowing to provide the desired flexibility and extendibility.

Additionally, even if both qualitative studies have been very fruitful, the size of the sample — 17 participants considering both studies — can be a threat to the validity of the results, even if detailed information about their design and execution have been provided and theoretical saturation has been achieved. However, qualitative studies are not meant to generate exhaustive and categorical results, and then, even if not all the possible concepts and relationships are represented in our model, we believe that they are still useful, as they are a good starting point. In this sense, we also considered if it would be necessary to have more details in the model — for example including all the possible subtypes of all the present concepts — or, alternatively, if it would be more suitable to provide a model with a higher level of abstraction. In order to maintain the validity of the results, we chose a simple inclusion, as already explained before. We consider that any concept mentioned by participants might be relevant to get insight into how CS researchers seek information, trying to keep an open mind to discover new concepts. That's why we use an inductive approach for the elaboration of our theory — in this case the conceptual model —, instead of trying to corroborate an initial hypothesis that might be dependent on the authors carrying out the study.

We recognize that using the mentioned inclusion/exclusion criterion can be too permissive as it may allow to take into consideration eccentricities from the participants of the qualitative study. Despite this, we have chosen to use it as it has allowed us to faithfully reflect the participants' opinions, reducing the coders influence to the maximum, and then increasing the validity and reproducibility of the results. Besides, we consider that the flexibility of the model ensures that, even if

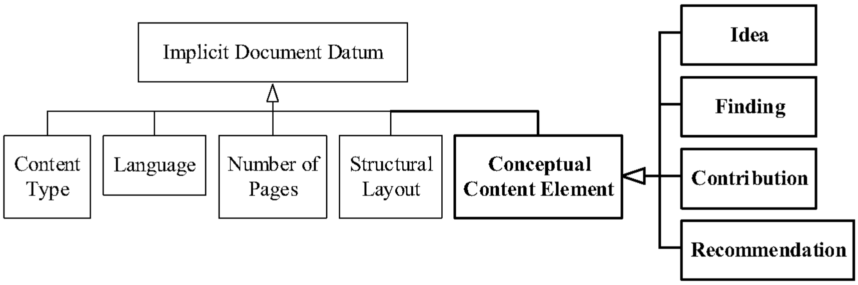

Fig. 8. Modification of part of the document's model to include a new type of concepts.

an eccentric specific type of concept is included in it, the latter can be perfectly ignored as the model is still valid and consistent. As an example, some readers may not consider a *webpage* as a type of document used in research, but ignoring or deleting this concept and all the ones depending on it does not affect the rest of the model.

The only aspects that have been mentioned by the participants but have not been included in the model are those related to the visualization and storage & management. In the case of the visualization, we haven't included any concept from those mentioned by the participants, as it is complex enough to be developed individually. Besides, we believe that the modeling of the visualization requires to be enriched with many other concepts, like those related to perception and visual theory. In the case of the storage & management tasks, we believe that it cannot be considered an information-seeking task — or at least not at the same level as filtering, exploring or reading — as its realization is usually independent of the information-seeking process. However, we consider both aspects essential to have a full picture of the information-seeking process, and that's why we are currently working on these subjects from the software development and human-computer interaction perspectives.

Nonetheless, even if we know that our model is not exhaustive, we claim its validity and usefulness as it is consistent, flexible and easily scalable, as using UML facilitates adding, modifying and deleting concepts and relationships without affecting its validity and consistency. As an example, the concept "*conference poster*" could be included in the model without any problem or difficulty, simply by adding it as a subtype of the document concept. Only by doing this, the concept "*conference poster*" would automatically inherit all the properties and relationships of the document concept, like having a set of information elements — authors' names, publication venue, publication date, sections, figures… —, a set of terms, and, if needed, an impact metric. A second example can also be provided to reflect how easy and safe it would be to include new concepts in the model: document's model does not include elements dealing with its semantic content like the ideas, findings, contributions — or recommendations it provides because this type of concepts were not explicitly related to the documents by the participants of the qualitative study. This is probably because these concepts are very abstract — and then difficult to recognize — and its recognition within the document requires an active analysis by the reader. This implies that it not obvious how these elements could be automatically extracted from the content of the document, like it can be done with its language or number of pages, for example. In any case, if it would be desired to complete the model with this aspects, the concept "*conceptual content element*" —that comprises the mentioned elements — could be considered as a new type of "*implicit document datum*", as reflected in Fig. 8.

Finally, we also acknowledge that our model is era-dependent, being as it is derived from the current experience of a set of CS researchers. We can foresee that some concepts included in the model might be not relevant in the future — like the ones related to manual access and manipulation of document collections — but, according to our inclusion/exclusion criterion, if at least one participant mentioned a concept, we have included it in the model. In this way, the model reflects the current situation for the participant CS researchers. Nevertheless, due to the flexible notation used, concepts can be deleted or substituted if they are deemed irrelevant. Again, the strength of the model is that, if the modification or addition is properly done, the rest of the model is not affected by the change, and the resulting model is still coherent and consistent.

12. Conclusion and future works

In this paper, for the first time a pictorial conceptual model has been presented in order to describe all the main concepts that are involved at some point in the activities that are undertaken by CS researchers in an information-seeking process. Two qualitative studies involving this kind of researchers with different levels of expertise have been performed in order to understand the aforementioned activities. Conceptual models, illustrated through a standard notation — UML — have been chosen, as they facilitate the understanding of the concepts and their relations, and they provide an unambiguous and graphical way to represent all the acquired knowledge.

Even if the information-seeking behavior has been extensively studied by other authors, all of them have adopted a narrowed approach leading to only understand the psychological aspect of the process, or its procedural steps. This results in studies focusing on some of the aspects intervening in the process, but leaving outside some others. Our proposal covers and defines every relevant aspect that may affect the seeking of information in the CS field for research purposes. To achieve this purpose, we propose to conceptually describe the information-seeking activities through a high level model — see Fig. 2 —, whose main components are later decomposed into more detailed conceptual sub-models:

- Sub-model 1: characteristics of the documents used as source of information during an information-seeking activity, both individually and as a collection;
- Sub-model 2: features of the tasks that can be performed in an information-seeking activity and how they are carried out;
- Sub-model 3: characteristics and specificities of CS researchers when seeking information;
- Sub-model 4: context in which an information-seeking activity takes place.

Using conceptual models and representing them through a visual notation has allowed us to make the difference with previous works, where models have been presented mainly in a written form or through very abstract and generic non-standard notations. Our model, on the contrary, describes for the first time information-seeking through a standard notation that ensures the understandability, transferability and unambiguity of the results. Apart from this, many authors have not to explained in detail the knowledge obtained during their analysis studies, and only provide some of the tips or guidelines they have obtained from them, directed to the design of software systems supporting information-seeking purposes. Due to the complexity and variability of the process, we consider that it is better to clearly reflect all the acquired knowledge as a framework instead of a list of recommendations. In this regard, our model has been conceived as a flexible, adaptable and extensible tool allowing to ignore, modify or incorporate new or field-dependent knowledge.

In summary, in this paper we propose an information-seeking model with the following features:

- Illustrated through a pictorial representation;
- Developed using a holistic approach that takes into consideration all the concepts and relationships intervening in information-seeking activities;
- Provides a general but comprehensive overview of the process
- Implemented using a standard, well-defined and non-ambiguous modeling language;
- Flexible enough to allow adapting it to concrete needs, like including new concepts that may appear in the future in order to reflect new realities, or to delete or modify existing concepts that may have become obsolete over time.

Further research would also be needed to extend the qualitative studies by interviewing a higher number of researchers as participants. This will allow us to validate the model, or to improve it by reflecting the new acquired knowledge. This, thus, would reduce the possible influence that individual biases may have introduced in the proposed model.

On the other side, due to the flexibility of the model and the number of similarities found with other models derived in different contexts of use, we believe that our model, even if stemming from a study focused on computer scientists, can be of great utility in other domains. However, we aim at verifying this hypothesis by using the obtained models as a framework to perform new studies. This will allow us to evaluate their validity and usefulness when modeling information-seeking activities in other research fields, or even in a general context.

Our long-term objective is to design and implement a software system based on the knowledge expressed by means of the proposed conceptual model, in order to provide full support to the research-oriented information-seeking activities, through the use of adaptive and customizable user interfaces, providing a high degree of usability, and generating very good user experiences. The model is going to be essential in the design process, as we need to identify not only the user's personal profile, but also his/her professional profile, that is, which is the information contained in the documents he/she usually manipulates, and how, when and for what purpose, what tasks he/she perform to fulfill each of these purposes, and in which context he/she usually seeks information. Having all this information updated dynamically and automatically will allow our system to "know" the user, and to anticipate to his/her needs and desires. In fact, nowadays, the model is being used as a scheme to implement an ontology capable to store all the concepts and relationships that have been identified. Then, this conceptual model may be used to design the storage system containing all the information related to the information-seeking activities, as the entity-relationship model is used to describe a given domain and guide the design of a classical relational database.

Acknowledgments

We thank the reviewers for their detailed comments and suggestions for improvement which have helped to greatly enhance the quality of the paper.

References

Anderson, C. J., Glassman, M., McAfee, R. B., & Pinelli, T. (2001). An investigation of factors affecting how engineers and scientists seek information. *Journal of Engineering and Technology Management, 18*(2), 131–155. http://doi.org/10.1016/S0923-4748(01)00032-7.

Association, A. P. (2001). *Publication manual of the american psychological association* (6th ed.). Washington, DC, USAWashington, DC: American Psychological Association.

Bates, M. J. (1989). The design of browsing and berrypicking techniques for the online search interface. *Online Review, 13*(5), 407–424. http://doi.org/10.1108/eb024320.

Benardou, A., Constantopoulos, P., Dallas, C., & Gavrilis, D. (2010). Understanding the information requirements of arts and humanities scholarship. *International Journal of Digital Curation, 5*(1), 18–33.

Boman, M., Bubenko, J. A., Jr., Johannesson, P., & Wangler, B. (1997). *Conceptual modelling*. Upper Saddle River, NJ, USA: Prentice-Hall, Inc.

Buckley, J., O'Brien, M. P., & Power, N. (2006). Empirically refining a model of programmers' information-seeking behavior during software maintenance. In *18th workshop of the psychology of programming interest group* (pp. 168–182). Retrieved from.

Byström, K., & Järvelin, K. (1995). Task complexity affects information seeking and use. *Information Processing & Management, 31*(2), 191–213. http://doi.org/10.1016/0306-4573(95)80035-R.

Case, D. O. (2012). *Looking for information: a survey of research on information seeking, needs and behavior*. Emerald Group Publishing Retrieved from *https://books.google.com [illegible] =YVX_RV7Wy9QC&pgis=1*

Cheuk, B. W., & Dervin, B. (1999). A qualitative sense-making study of the information-seeking situations faced by professionals in three workplace contexts. *The Electronic Journal of Communication, 9*(2). Retrieved from http://www.cios.org/EJCPUBLIC/009/2/009215.html.

Corbin, J., & Strauss, A. (2014). *Basics of qualitative research: techniques and procedures for developing grounded theory*. Sage Publications.

Dervin, B. (1992). From the mind's eye of the user: The sense-making qualitative-quantitative methodology. *Qualitative Research in Information Management, 9*, 61–84.

Dervin, B. (2003). An Overview of Sensemaking research: Concepts, Methods, and Results to Date. In B. Dervin, L. Foreman-Wernet, & E. Lauterbach (Eds.), *Sense-Making Methodology Reader: Selected Writings of Brenda Dervin*. Creskill, NJ: Hampton Press http://www.ideals.illinois.edu/bitstream/handle/2142/2281/Dervin83a.htm.

Dervin, B., & Nilan, M. (1986). Information needs and uses. *Annual Review of Information Science and Technology, 21*, 3–33.

DiCicco-Bloom, B., & Crabtree, B. F. (2006). The qualitative research interview. *Medical Education, 40*(4), 314–321. http://doi.org/10.1111/j.1365-2929.2006.02418.x.

Domik, G. O., & Gutkauf, B. (1994). User modeling for adaptive visualization systems. In *Visualization, 1994., visualization '94, Proceedings., IEEE conference on* (pp. 217–223). CP24.

Ellis, D. (1989a). A behavioural approach to information retrieval system design. *Journal of Documentation, 45*(3), 171–212. http://doi.org/10.1108/eb026843.

Ellis, D. (1989b). A behavioural approach to information retrieval system design. *Journal of Documentation, 45*(3), 171–212. http://doi.org/10.1108/eb026843.

Ellis, D. (1993). Modeling the information-seeking patterns of academic researchers: A grounded theory approach. *The Library Quarterly, 63*(4), 469–486. http://doi.org/10.2307/4308867.

Ellis, D., Cox, D., & Hall, K. (1993). A comparison of the information seeking patterns of researchers in the physical and social sciences. *Journal of Documentation, 49*(4), 356–369. http://doi.org/10.1108/eb026919.

Ellis, D., & Haugan, M. (1997). Modelling the information seeking patterns of engineers and research scientists in an industrial environment. *Journal of Documentation, 53*(4), 384–403. http://doi.org/10.1108/EUM0000000007204.

Feder, A.. *Your BibTeX resource* Retrieved April 1, 2016, from http://www.bibtex.org/.

Ford, N.. *Introduction to information behaviour* https://books.google.es/books/about/Introduction_to_Information_Behaviour.html?id=5Q64MQEACAAJ&pgis=1.

Foster, A. (2004). A nonlinear model of information-seeking behavior. *Journal of the American Society for Information Science and Technology, 55*(3), 228–237.

Freund, L., Toms, E. G., & Waterhouse, J. (2006). Modeling the information behaviour of software engineers using a work - task framework. *Proceedings of the American Society for Information Science and Technology, 42*(1). http://doi.org/10.1002/meet.14504201181.

George, C. A., Bright, A., Hurlbert, T., Linke, E. C., St Clair, G., & Stein, J. (2006). Scholarly use of information: Graduate students' information seeking behaviour. *Information Research, 11*(4) paper 272.

Glaser, B. G. (1978). *Theoretical sensitivity: Advances in the methodology of grounded theory*. Sociology Press.

Glaser, B. G., & Strauss, A. L. (2009). *The discovery of grounded theory: Strategies for qualitative research*. Transaction Publishers.

Godbold, N. (2006). Beyond information seeking: A general model of information behaviour. *Information Research, 11*(4). Retrieved from http://www.informationr.net/ir/11-4/paper269.html.

Gralewska-Vickery, A. (1976). Communication and information needs of earth science engineers. *Information Processing & Management, 12*(4), 251–282. http://doi.org/http://dx.doi.org/10.1016/0306-4573(76)90065-0.

Hertzum, M., & Pejtersen, A. M. (2000). The information-seeking practices of engineers: Searching for documents as well as for people. *Information Processing & Management, 36*(5), 761–778. http://doi.org/http://dx.doi.org/10.1016/S0306-4573(00)00011-X.

Järvelin, K., & Wilson, T. D. (2003). On conceptual models for information seeking and retrieval research. *Information Research, 9*(1), 163. Retrieved from http://www.informationr.net/ir/9-1/paper163.html.

Krikelas, J. (1982). Information-seeking behavior: Patterns and concepts. *Drexel Library Quarterly, 19*(2), 5–20. Retrieved from http://eric.ed.gov/?id=EJ298483.

Krueger, R. A., & Casey, M. A. (2015). *Focus groups: A practical guide for applied research* (5th ed.). Thousand Oaks, California: SAGE Publications.

Kuhlthau, C. C. (1991). Inside the search process: Information seeking from the user's perspective. *Journal of the American Society for Information Science, 42*(5), 361–371. http://doi.org/10.1002/(SICI)1097-4571(199106)42:5⟨361::AID-ASI6⟩3.0.CO;2-#.

Kwasitsu, L. (2003). Information-seeking behavior of design, process, and manufacturing engineers. *Library & Information Science Research, 25*(4), 459–476. http://doi.org/10.1016/S0740-8188(03)00054-9.

Leckie, G. J., Pettigrew, K. E., & Sylvain, C. (1996). Modeling the information seeking of professionals: A general model derived from research on engineers, health care professionals, and lawyers. *The Library Quarterly, 66*(2), 161–193.

Makri, S., Blandford, A., & Cox, A. L. (2008). Investigating the information-seeking behaviour of academic lawyers: From Ellis's model to design. *Information Processing & Management, 44*(2), 613–634. http://doi.org/10.1016/j.ipm.2007.05.001.

Manning, C. D., Raghavan, P., & Schütze, H. (2008). *Introduction to information retrieval*. Cambridge University Press.

Marchionini, G. (1997). *Information seeking in electronic environments*. Cambridge University Press.

Marchionini, G., & White, R. (2007). Find what you need, understand what you find. *International Journal of Human–Computer Interaction, 23*(3), 205–237. doi:10.1080/10447310701702352.

Marsh, T. (2003). Staying There: An Activity-Based Approach to Narrative Design and Evaluation as an Antidote to Virtual Corpsing. In G. Riva, F. Davide, & W. A. IJsselsteijn (Eds.), *Being There: Concepts, effects and measurements of user presence in synthetic environments* (pp. 85–96). Amsterdam, The Netherlands: IOS Press. Retrieved from http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.4.2518.

Meho, L. I., & Tibbo, H. R. (2003). Modeling the information-seeking behavior of social scientists: Ellis's study revisited. *Journal of the American Society for Information Science and Technology, 54*(6), 570–587. http://doi.org/10.1002/asi.10244.

Moral, C., de Antonio, A., Ferre, X., & Lara, G. (2015). A coding system for qualitative studies of the information-seeking process in computer science research. *Information Research, 20*(4). Retrieved from http://www.informationr.net/ir/20-4/paper699.html#.VpfgN5PhC-p.

Niedźwiedzka, B. (2003). A proposed general model of information behaviour. *Information Research, 9*(1). paper 164. Retrieved from http://www.informationr.net/ir/9-1/paper164.html.

O'Brien, M. P., & Buckley, J. (2005). Modelling the information-seeking behaviour of programmers - An empirical approach. In *13th international workshop on program comprehension (IWPC'05)* (pp. 125–134). IEEE.

Object Management Group (2016). *Unified modeling language (UML)*. Retrieved June 8, 2016, from *http://www.omg.org/spec/UML/*.

Onggo, S. (2010). Methods for Conceptual Model Representation. In S. Robinson, R. Brooks, K. Kotiadis, & D.-J. Van Der Zee (Eds.), *Conceptual modelling for discrete-event simulation* (pp. 337–354). CRC Press.

Holland, M. P., & Powell, C. K. (1995). A longitudinal survey of the information seeking and use habits of some engineers. *College & Research Libraries, 56*(1), 7–15. Retrieved from http://cat.inist.fr/?aModele=afficheN&cpsidt=3015364.

Pinelli, T. E. (1991). The information-seeking habits and practices of engineers. *Science & Technology Libraries, 11*(3), 5–25. http://doi.org/10.1300/J122v11n03_02.

Preece, J., Rogers, Y., & Sharp, H. (2015). *Interaction Design: beyond human-computer interaction* (4th ed.). Wiley.

Richter, H., & Marz, L. (2000). Toward a standard process: The use of UML for designing simulation models. In J. A. Joines, R. R. Barton, K. Kang, & P. A. Fishwick (Eds.), *Proceedings of the 2000 winter simulation conference. Orlando, FL, USA: vol. 1* (pp. 394–398).
Saldaña, J. (2015). *The coding manual for qualitative researchers*. Sage.
Stevens, P., & Pooley, R. (2006). *Using UML: software engineering with objects and components* (2nd ed.). Essex, England: Pearson Education.
US National Library of Medicine (2012). *NLM journal archiving and interchange tag suite* Retrieved April 1, 2016, from *http://dtd.nlm.nih.gov/*.
University of Chicago (2010). In University of Chicago Press Staff (Ed.), *The Chicago manual of style* (16th ed.). Chicago, USA: University of Chicago Press Retrieved from *http://www.chicagomanu*
Wilson, T. D. (1981). On user studies and information needs. *Journal of Documentation, 37*(1), 3–15. http://doi.org/10.1108/eb026702.
Wilson, T. D. (1997). Information behaviour: An interdisciplinary perspective. *Information Processing & Management, 33*(4), 551–572. http://doi.org/http://dx.doi.org/10.1016/S0306-4573(97)00028-9.
Wilson, T. D. (1999a). Exploring models of information behaviour: The "uncertainty"project. *Information Processing & Management, 35*(6), 839–849.
Wilson, T. D. (1999b). Models in information behaviour research. *Journal of Documentation, 55*(3), 249–270. http://doi.org/10.1108/EUM0000000007145.