Chapter 4

Information Visualization Model

In the last decades, processing large amount of data has become very popular and common in multiple fields —like economy, health care, sociology, information retrieval, etc. In fact, big data and data mining have suffered a great growth in the last years, due to the great increase of data that are obtained from the huge amount of devices and digital services used today —for example smartphones and social network applications—, but also because of the rapid increase of processing capabilities of computers. In this sense, it is essential not only to process this huge amount of data, but also to visualize it in a proper manner so that it becomes understandable for and processable by the user. As Shneiderman stated in one of the reference papers in information visualization, "Abstract information visualization has the power to reveal patterns, clusters, gaps, or outliers in statistical data, stock-market trades, computer directories, or document collections" (Shneiderman, 1996, p. 2).

However, information visualization is not only aimed at processing statistical quantitative data and recognizing patterns and tendencies. In the case of information retrieval, where users usually have a concrete information-seeking purpose to be achieved, information visualization is an essential step to help users analyzing hundreds or thousands of documents in order to find those that fulfill their information needs. In fact, even if the data processing is extremely good and the information system is able to analyze very well the content of the documents and how they relate, if the presentation and manipulation of these results is not properly undertaken, the user may find difficult, inefficient and unappealing to use the said system. This is due to the inner perception and cognitive capacities of humans to recognize, process, and analyze visual representation in order to infer complex information and resolve logical problems (Bertin, 1981). Then, the main characteristics of information visualization, regardless its purpose, is to "use [...] computer-supported, interactive, visual representations of abstract data to amplify cognition" (Card, Mackinlay, & Shneiderman, 1999, p. 7).
In order to properly conceptualize how information has to be visualized to facilitate the accomplishment of the users’ information-seeking purposes, it is necessary to identify which are the main aspects that have to be taken into account when designing an information visualization system.

During the last years, many information visualization researchers have claimed that visualization cannot be considered as a context-independent activity as its efficiency and utility directly depend on how well the visualization and interaction techniques offered by the system actually facilitate the user carrying out a given information-seeking task, according to his/her preferences, characteristics and purposes, and considering the context in which the information visualization activity is being performed (Einsfeld et al., 2006). As seeking information is a complex and heterogeneous process that depends on many factors, it seems reasonable to think that an information visualization system should provide different visualization and interaction techniques, so that it can cover different user profiles, different information needs and different contexts of use. In the proposed model is represented this need of variety through the visualization workspace concept. As an example, an information visualization system could offer two very different workspaces: one of them would display information-seeking results as an ordered list, while the other would display the same items through a graph (Eades & Feng, 1996; Eades & Huang, 2000; Fisher, Dhamija, & Hearst, 2001) or a hierarchical tree (G. G. Robertson, Mackinlay, & Card, 1991; Shneiderman, 1992; Plaisant, Grosjean, & Bederson, 2002; B. Johnson & Shneiderman, 1991). In this case, the user could choose which one the visualizations to employ or even visualize both of them simultaneously, depending on his/her needs.

Because of this, and as it has been done for the information-seeking process model —see Chapter 3—, in this chapter these aspects have been formalized through a UML-based conceptual model that is comprehensive, flexible and adaptable enough to represent the concepts and relationships intervening in the information visualization process in a research-oriented information-seeking context. The full model is included in Appendix C.4

In order to facilitate the understanding of the model, examples of real information visualization systems of different types are provided along the entire chapter, namely the very well-known Google Scholar\(^1\), Calimaco (de Antonio et al., 2012, 2013a, 2013b) —that has been briefly described in Chapter 1—, and 3D Explorer (Perez & de Antonio, 2003, 2004), that makes use of a multi-axis system to spatially organize the documents’s representations according to the topic(s) they address —each axis represents a topic—, and that also uses the color to illustrate the thematic similarity between documents.

\(^1\)https://scholar.google.es/
3D Explorer, additionally, allows displaying the documents as a colored alphanumerical list.

### 4.1 Information visualization activity

An information visualization activity can be defined as a specific use of an information visualization system, as reflected in Figure 4.1. This use is unique as it takes place in a given context, and it is performed by a given user with a certain mood, characteristics and with a given information need, who chooses some of the visualization workspaces available in the system to fulfill his/her information-seeking purpose. This election establishes which are the input and output devices that have to be used by the user, which type of visual representations are going to be used to display the information, and which information visualization task is the user going to be able to perform to meet his/her target.

A device can simultaneously display more than one visualization workspace, for example by splitting its screen into different parts, one per visualization workspace. Similarly, the model also supports defining information visualization activities that are performed in a multi-device setting, where visualization workspaces are displayed in different devices. An information visualization activity, then, could be carried out at the same time, for example, in two visualization workspaces, one of them displayed in a tablet and used to represent a corpus of documents in a given way, while the other one could be held in a computer screen and could display the same set of documents but in a different way. In this case, manipulating information in any of the visualization workspaces could cause the same effect in the other visualization workspace in order to maintain the consistency. Then, for instance, if a user selects three document titles from the list in one of the visualization workspaces and deletes them, the visual metaphors used to represent these documents in the other visualization workspace have to automatically disappear.

![Figure 4.1 Information visualization activity](image-url)
Then a visualization workspace can be defined as a digital environment where an information visualization activity is performed. In this context, one of the interesting aspects of digital environments is that they can be infinite. This is especially relevant in information visualization as usually the amount of information to be displayed is huge, and then there is theoretically no problem to allocate it all. The problem, then, lies with the screen of the output device, as it has a limited size and then, when the information occupies too much space in the digital environment, it cannot illustrate it all at the same time—or at least not in a way that it is visible and processable by the user. To solve these problems, it is required to allow the user to navigate through the environment by manipulating the view from which he/she observes the visualization workspace—see Section 4.5. This means that, at a given moment, it is possible that not all the visual representations that are present in the visualization workspace are visible for the user, even if they actually exist in the environment. In the proposed model, this portion of the visualization workspace that is visible by the user is represented by the view concept. In Google Scholar, for example, the visualization of the collection of documents returned as result of the raised query is performed in a unique visualization workspace. Nevertheless, at every moment, the user can only visualize this workspace through a static view, where only ten or twenty documents—depending on the user configuration—are visible and manipulable. Then, if the user wants to see the rest of the documents, he/she has to change the view. In 3D Explorer, on the contrary, two workspaces are offered. One of them—see Figure 4.2—is a three-dimensional environment where a set of concurrent lines, all stemming from the same origin, illustrate the clusters into which the documents are classified according to the topic they deal with. The second information visualization workspace represents the same documents, also clustered by the topic they address, but as a one dimensional textual list—see Figure 4.3. In this case, color is again used to illustrate the thematic clustering—see Figure 4.3.a—and the color hue is used to illustrate that the documents can be categorized in more than one thematic cluster—see Figure 4.3.b. Additionally, documents representations are sorted based on two criteria: first of all, each document is inserted in a a drop-down item that represents the topic—see Figure 4.3.c--; and secondly, within each of these drop-downs lists, the position of the alphanumerical documents representation is modified in order to sort them alphabetically—see Figure 4.3.d—

So, each visualization workspace has to be considered as an independent digital environment that has its own features. Among these features, it can be defined when a visualization workspace has to be visible and when not. This decision may be taken by the user but, occasionally, the information visualization system can automatically
4.1 Information visualization activity

(a) Control widget represented through a metaphor; (b) Control widgets represented through GUI components (buttons and radio buttons) with an associated alphanumerical representation; (c) Modification of visualization dimension “color” to illustrate the thematic clustering; (d) Document’s filename and keywords represented through an alphanumerical representation; (e) Modification of the visualization dimension “color hue” to illustrate that the document is clustered in two groups; (f) Graphical representation of the query.

Figure 4.2 Example of information seeking and visualization system: 3D Explorer (First workspace).

display, or at least suggest to display, a given visualization workspace depending on the context in which the visualization is being performed, on the information visualization task the user is currently performing or on the user preferences. Then, either the system has to be able to recognize when the user may benefit from using a given visualization workspace, or the user has to be able to indicate the system when he/she wants a visualization workspace to get visible, or invisible — in the model, this concept is reflected as the visibility criterion. As an example, both Google Scholar and Calimaco
Thematic clustering illustrate by changing the visualization dimension "color"; (b) Modification of the visualization dimension "color hue" to reflect that a document can be categorized in more than one cluster; (c) Sorting based on the cluster in which are categorized the documents (affect the visibility state of their representations); (d) Modification of the visualization dimension "spatial position coordinate" (y axis) to illustrate that documents' representation are organized alphabetically.

Figure 4.3 Example of information seeking and visualization system: 3D Explorer (Second workspace).

have a unique visualization workspace, and therefore there is no visibility criterion—the workspace is always visible. In the case of 3D Explorer, the three-dimensional workspace is also always visible, while the one-dimensional workspace is only visible on-demand.

Apart from these aspects, there are others features that specifically depend on the aim of the information visualization system and on the specific design of its visualization workspaces, as for example the interaction style it follows, or its dimensionality, that is the number of dimensions it uses to present the information. In this work have been
considered the interaction styles proposed by Shneiderman, Plaisant, Cohen, and Jacobs (2009), but new ones can be added if required:

- **Menus, Forms and Dialog Boxes.** This is the interaction performed in classical systems with GUI components like buttons, drop-down menus, textboxes, and so on. Google Scholar —see Figure 4.13— is an example of this type of interaction style.

- **Commands.** The user interacts with the systems by providing textual, vocal or gestural orders. The classical command line environment, or the more recent vocal assistants —like Apple Siri and Microsoft Cortana— are examples of it.

- **Direct Manipulation.** The user interacts with the system by directly manipulating visual metaphors, like the ones used in Calimaco —see Figure 4.6— or in 3D Explorer —see Figure 4.2.

With respect to the dimensionality, a visualization workspace can be:

- **One-dimensional,** when information can only be located along one axis. Visualizations that illustrate the information as a list of items, like Google Scholar, are an example of one-dimensional workspaces;

- **Two-dimensional,** if the information is distributed along both horizontal and vertical axes but depth is not considered, like in FilmFinder (Ahlberg & Shneiderman, 1994) or in GRIDL (Shneiderman, Feldman, Rose, & Ferre, 2000); and

- **Three-dimensional,** if information is positioned in a 3D space, like Calimaco or 3D Explorer.

Most likely, there exist many other concepts that could define a visualization workspace. However, it is not intended to represent them all, but only those that are considered quite common to any information visualization system, as the present model is flexible and expandable enough to be adapted or augmented, if needed.

## 4.2 Device

First of all, it is necessary to define which kind of devices are going to be supported by the information visualization system, as the designer has not only to consider how the information is going to be visualized, but also how it is going to be manipulated. This is essential as it narrows the type of devices that are supported by the information
visualization system, taking into account all the characteristics that are necessary to correctly visualize the information environment and to carry out all the information visualization tasks offered by the system. As an example, if the visualization is going to be performed in an immersive three-dimensional environment, then the device displaying it has to be able to produce stereoscopic images. Analogously, if one of the information visualization tasks, like selecting some items from the information-seeking results, requires the user performing in-air gestures, then a 6 degrees-of-freedom input device is required. Figure 4.4 illustrates some of the device features that have to be taken into account to describe how the user visualizes and interacts with the information.

Output devices’ function is to transmit the information provided by the system to the user by the use of digital means —for example a computer monitor, a projector or a head-mounted display—, whereas input devices allow the user to give orders and information to the system so it can process and execute them —for example a keyboard, a microphone or an optical tracking system. Additionally, with the appearance of touch-based technologies, there are input/output devices that allow transmitting information in both ways, like smartphones, tablets, touch tables or tangible devices.

Each of these devices has associated a set of features that allow describing their functionality, their capabilities, and their restrictions. Some of these features can characterize any device, regardless its type, like transportability, while others may apply only to input or output devices:

- **Transportability.** It has to be considered if the device can be carried, and then the system can be used under different environmental conditions, or if, on the contrary, the device has a permanent location. In the case of input devices, this can influence on how the user interacts with the system because of physical restrictions. As an example, it is not the same to interact with a system in an
office, where the user is sitting and the environmental conditions are stable, than to manipulate a system held in a smartphone while traveling by subway, where user’s position may vary —he/she may be sitting or standing— and he/she may have more or less freedom of movement depending on how busy is the subway car.

- **Input Method.** Using classical input methods, like a keyboard and a mouse, is very different, for example, from inputting data through gestures or voice commands, and then specific interaction techniques have to be provided to allow the user to perform the information visualization task in the most intuitive, comfortable, efficient and usable way. The selection of the input method is closely related with the type of data that has to be provided to the system by the user.

- **Degrees of Freedom.** It is essential to know how many degrees of freedom the device is allowed to manipulate, as this will directly affect which kind of interaction techniques can be performed in the visualization workspace. As an example, it is not the same to navigate through a three-dimensional virtual environment using a device that only can manipulate two degrees-of-freedom, like a classical mouse, than using a six degrees-of-freedom device, like a 3D mouse. In the first case, the degrees-of-freedom that are not directly manipulable with the device have to be manipulated through software-based solutions —like 3D widgets allowing to translate and rotate the view in any axis—, while in the second case, both translation and rotation can be performed directly by manipulating the device. This aspect is closely related to the dimensionality of the workspace.

- **Screen Definition.** As visual features, like color and texture, are going to be used, it is essential to know which is the fidelity level of the device when displaying these features. In fact, if the definition is not good enough, there is a risk that some of the values of some visual features may not be displayed as expected, which would imply the loss or misunderstanding of the information carried by these visual features.

- **Screen size.** It has to be considered when designing the information visualization system and its workspaces as it directly affects how much information can be displayed in a single view, or how many visualization workspaces can be simultaneously displayed.

As stated before for the visualization workspace features, it is not pretended to exhaustively represent all the features that can define a device, but only to point out those concepts that are considered relevant and general, and thus should be addressed and considered when designing an information visualization system. A more complete
and exhaustive modeling of devices and their interactions has been performed by Jeronimo, de Antonio, Mendez, and Ramirez (2009).

4.3 Visual representations

In terms of visualization, one of the first design decisions that has to be taken is how the information is going to be displayed. Usually, in classical information systems results are displayed through a list with more or less details. In the literature, nonetheless, can also be found many examples of systems that use a graphical visualization to display the results of an information-seeking activity (K. M. Fairchild, Poltrock, & Furnas, 1988; Card, Robertson, & Mackinlay, 1991; G. G. Robertson et al., 1991; J. D. Mackinlay, Robertson, & Card, 1991; Rekimoto & Green, 1993; Spoerri, 1993; Ahlberg & Shneiderman, 1994; Benford & Mariani, 1995; Benford et al., 1995; Risch et al., 1997; Shneiderman et al., 2000; Zhang, 2001; Perez & de Antonio, 2004; Zhang & Nguyen, 2005; Einsfeld et al., 2006; Ahn & Brusilovsky, 2013; de Antonio et al., 2013b).

As mentioned in Chapter 3, in the research domain, the information-seeking process mainly revolves around one entity: the document. That’s why in the proposed model—see Figure 4.5— it is considered that document representation is the core of the visualization, as it conveys information not only about the document’s content, but also about its relationships with other documents. Besides this, in the present model is also considered the possibility of needing to represent a conglomerate of documents as a single entity, as CS researchers typically need to manage many very large collections of documents, whether from own or outside sources, and then they may find useful to represent the whole collection as a single element.

![Visual representations of the elements displayed in the visualization](image)

Figure 4.5 Visual representations of the elements displayed in the visualization

However, beyond documents, there are other aspects that need to be represented in a visualization workspace. When visualizing information, users have to be able to perform
some information visualization tasks in order to navigate through the information, and
to manipulate it to satisfy their information needs. These tasks have to be triggered
somehow, and, depending on the interaction style used to design the visualization
workspace and on the input method of the input device, it may be needed to provide
some software-based triggers to launch the desired tasks. Obviously, these triggers
are also used to perform tasks that are not directly related with the visualization,
like saving a given visualization, or selecting the source from where are obtained the
documents —Google Scholar, for example, allows the user to define if the search will be
performed in the whole web, or in the personal collection. In the proposed model these
elements are called control widgets. Google Scholar provides a set of filtering, sorting
and configuration options through selectable textual elements, check-boxes, buttons and
visual metaphors —see Figures 4.13(c) and 4.13(f) to 4.13(j) —, while Calimaco offers
the user eight visual metaphors to perform different information visualization tasks
—see Figure 4.6(a). 3D Explorer, meanwhile, offers two types of control widgets. The
first one allows rotating the view around the vertical and horizontal axes, and to zoom
in and out by the use of a set of arrows —see Figure 4.2(a). Additionally, it also allows
to rotate the view by using the camera as reference point —instead of the axes—, by
selecting the "Look at" mode in component of the GUI. This component makes part of
a set of radio-buttons, check-boxes and buttons —see Figure 4.2(b)— that are provided
in order to launch some functionalities, like changing the rotation mode, making the
view rotate automatically, making visible the second visualization workspace —called
"Tree view"—, modifying the number of topics into which the documents are clustered,
representing a query in the information space as a new virtual document, or modifying
the view options.

In the same way, information visualization systems have to provide the user, either
on-demand or according to the context, some sort of feedback about how he/she has to
interpret what he/she is seeing. This legend is especially useful when the workspace
makes use of various visual features to convey some kind of information about the
documents, or to explain the user the meaning of some graphical representations,
especially if they are abstract. It is required, then, to illustrate somehow not only the
visual feature and/or the graphical representation that needs to be explained, but also
a textual label providing the explanation. For instance, the system could use a star
as a metaphor to illustrate that a document has been marked as favorite by the user,
but it could also draw a pin near the document that has been blocked by the user in
order to avoid it being affected by filtering tasks. The first metaphor, even if abstract,
could be understood because it has become a social standard due to its use in many
systems —mainly Internet browsers—, while the user could associate the second one to
the real object and infer its meaning. However, in both cases, it cannot be assured that every user would intuitively recognize both metaphors, and then it would be probably needed to explain them in the legend. In order to explain the values of a visual feature, it could be also possible to make use of the label by assigning to it the same value in such visual feature—if, for instance, red color illustrates documents published before 2000 and green color the rest of documents, a label with the text "Before 2000" written in red, and another label with the text "After 2000" written in green would perfectly explain the use of the color in the visualization.

Besides legend elements, there are other visual representations that can have an associated label. A document collection representation, for instance, can have a label attached in order to indicate the number of document representations that are included in it; it can also have attached a label with the most frequent term appearing in the document, like it is done in Calimaco—see Figure 4.6(b)—; or a control widget representation can have associated a label indicating the name of the functionality or information visualization task it triggers, as it is done in 3D Explorer—see Figure 4.2(b). In order to conceptualize these labels, and any other representation that may be needed to illustrate, directly or indirectly, something about the documents, the present model includes the concept other visual representations. In the case of Calimaco, these kind of representations are used to illustrate the connecting lines—see Figure 4.7(b)—
4.3 Visual representations

stemming from the cursor —represented as two fingerprints, as reflected in Figure 4.7(a)— that indicate which document representations are close to it. Additionally, if a document representation is selectable, then the line connecting it to the cursor becomes bold and dashed —see Figure 4.7(c).

Regardless the concept they represent, visual representations have a set of features that are related to their visibility —as a visualization workspace, a visual representation has a visibility criterion and a visibility state—, and to their visual appearance —each visual representation has a certain amount of visual features defined. These features are illustrated in Figure 4.8.

Among these features —in the present model they are called visualization dimensions—, there may be some that can be modified by the system in order to provide some kind of information. Conceptualizing these dimensions is essential, as the information to be visualized is often very complex, and then it requires using more than one aspect of the visualization. Shneiderman (1996) proposes a type by task taxonomy that allows

Figure 4.7 Example of use of other visual representations

(a) 3D fingerprints representing the cursor; (b) Lines illustrating which document representations are close to the cursor; (C) Bold and dotted line illustrating that the document connected to the cursor is close enough to be selected.
classifying the data that can be visualized in an information visualization system into seven categories: one-, two-, three- and multi-dimensional data, temporal data, and tree and network data. This research deals with multi-dimensional and network data, as documents and document collections aggregate a set of data, and many of these data allow linking documents and document collections —see Section 3.3 in Chapter 3— that, depending on the purpose of the information-seeking task and on the user’s preferences and characteristics, may have to be displayed somehow in the visualization. And this is precisely the aim of visualization dimensions. In Google Scholar, documents are represented by text entries, each of them providing different types of information —see Figure 4.13(d). In fact, each of these chunks of information is formatted in a different way in order to facilitate their visual recognition by the users. To be more concrete, color is used to convey the type of the information: the title is written in blue, the authors in green, and the abstract in black. In the case of Calimaco and 3D Explorer, color is used to represent the document’s membership to a cluster —see Figures 4.6(c) and 4.2(c).

As reflected in Figure 4.9, in the proposed model it is considered that a given visual representation can be illustrated with different modalities that can be categorized into two main groups: alphanumerical and graphical representations. The first case covers the use of any kind of textual representation to transmit the information, like the one used in Google Scholar to represent the documents, or the labels used to illustrate the most frequent term of a document in Calimaco —see Figure 4.6(b)— and of a topic in 3D Explorer —see Figure 4.2(d). Graphical representation, on the contrary, groups all the geometrical, pictorial, metaphorical, and non-textual representations. This includes the GUI Components that are typically used in software systems —menus, buttons, sliders, check-boxes, and so on—, the visual metaphors that illustrate, more or less reliably, real concepts —like a star, a drawing pin, or the bin icon—, and the geometrical figures, that can convey some information in some contexts —for instance, a
circle or a square that enclose a group of documents representations probably illustrates that they are related somehow, in the same way it does a line linking two document representations.

4.4 Visualization dimensions

Then, a visualization dimension can be defined as a visual representation feature whose value can be modified in order to provide some kind of information about the represented document, document collection, legend element, etc. These visualization dimensions can be categorized into two groups, as reflected in Figure 4.10, depending on whether their values can be sorted based on an ordering criterion —ordinal dimensions—, or not —nominal dimensions.

Some examples of ordinal dimension are:
• **Spatial position.** Within the visualization workspace, a visual representation can be located in an infinity of positions. Each of these positions is determined by one to three spatial position coordinates (axes x, y, and/or z), depending on the dimensionality of the visualization workspace. Focusing on each of these three axes, a visual representation can be above or below other visual representations, but it also can be rightmost or leftmost, or closer or farther. For example, in Google Scholar documents are presented in an ordered list based on their relevance with respect to the query—see Figure 4.13(b)—. Then, the vertical spatial position coordinate of a document representation depends on its relevance. In Calimaco, on the contrary, document representations are located in a three-dimensional space, according to their topic, so that documents dealing with similar or related topics are represented in close positions in the visualization workspace. Similarly, in 3D Explorer the location of the document representations depends on the topics they deal with, but in this case the spatial position is figured out with respect to the lines that represent the different clusters. Then, the closer a document is represented to a given cluster line, the more its content is related with the topics represented by the cluster.

• **Color.** In its basic form, colors can be used as a set of unordered and unrelated values—as an example, yellow, green and blue cannot be ordered under any criterion. However, color admits different levels of detail, and depending on how it is used and manipulated, it can be also considered as an ordinal dimension. For this, it has to be divided into a set of ordinal dimensions, like those proposed by Joblove and Greenberg (1978) in the HSL model, where colors into are decomposed into 3 aspects—see Figure 4.11—whose values can inherently be sorted according to an ordering criterion:

  – **Color saturation.** The saturation of a color defines how vivid it is. Then, an intense color is more saturated than a dull color.

  – **Color hue.** The hue of a color defines *the degree to which a stimulus can be described as similar to or different from stimuli that are described as red, green, blue, and yellow* (M. D. Fairchild, 2004). Then, it reflects how similar a given color is to a pure color. As an example, in 3D Explorer the color hue of a document representation varies according to the degree of membership of the document to a given cluster. Figure 4.2.e illustrates that
both documents are related with the cluster represented with a light blue line, but one of them is also slightly related with another cluster. To reflect this in the visualization, the color hue of this document representation is modified so that the user can visually identify that the color is similar to the cluster line one, but less pure.

– Color lightness. The lightness of a color defines how bright it is. Then, a color can be much more illuminated than other colors that are darker.

• Size or Thickness. A visual representation can be bigger/thicker or smaller/thinner than other visual representation. As an example, in Google Scholar the size of the font used in the title of a document is slightly bigger than the rest of the information displayed —see Figure 4.13(e). In Calimaco and 3D Explorer, depending on the level of zoom of the view, document representations are bigger —if the camera has zoomed in— or smaller —if the camera has zoomed out. Additionally, in Calimaco only document representations that are close enough to the camera have their associated label —with the most frequent term— displayed.

On the other side, some of the nominal dimensions that can be used to represent a visual representation —besides the color in its basic form— are:

• Shape. A visual representation can be, for example, a circle, a sphere, a square, a triangle or a hexagon, but also a house, a drawing pin, or an arrow.

• Texture. As background, a visual representation can have some kind of pattern, like parallel lines —that can be drawn vertically, horizontally, or in diagonal—, grids, or regularly located geometric figures —like points, squares, diamonds, etc.

• Highlighting. Each visual representation can have associated a binary state: shining or not. For example, a bright light surrounding a visual representation can be used to distinguish the representations that have been selected from those that have not.

• Movement. At a basic level, a visual representation can rotate, jump or vibrate, for example, to illustrate a binary value, like the highlighting. Additionally, if the movement speed can be modified, it can reflect multivalued states —in this case, speed can even be considered an ordinal visualization dimension.

• Font Properties. In the particular case of alphanumerical representations, there are a number of properties related to the text font used that can be modified, like the type of font —for example Arial or Times New Roman—, or its format
—serif or sans serif, bold, underlined, italics, etc. In this case, the font size is not included in this category, as its values can be sorted, and then it’s an ordinal dimension.

Many of these visualization dimensions, have been already identified as visual features that can be effectively used in information visualization, as they are considered retinal properties that facilitate the preattentive processing of what the user is seeing. Using these kind of properties is a key factor in information visualization as it allows to perform an automatic processing, which is "superficial, parallel, can be processed non-foveally, has high capacity, is fast, cannot be inhibited, is independent of load, unconscious, and characterized by targets 'popping out' during search" (Card et al., 1999, p 25). Among the visualization dimensions existing in the literature that are considered preattentive, there are the spatial position, the color —both individually, and decomposing it into hue, saturation and value, which is a model similar to the HSL model used here—, the texture, the shape, the size, the direction of motion, and the flickering (Bertin, 1981; J. Mackinlay, 1986; MacEachren, 1995; Healey, Booth, & Enns, 1996; Card & Mackinlay, 1997).

The specific type of visualization dimension to be used depends on the kind of information that has to be displayed and on the aim of the visualization. In fact, there are three types of data that can be displayed in an information visualization system: nominal, ordered and quantitative (Card & Mackinlay, 1997). So, in order to faithfully transmit all the information provided by these data, it is required to make use of visual features allowing to transmit their specifications, like the possibility of being ordered. Then, if the aim is, for example, to cluster a set of document representations according to a given criterion, then using either a nominal dimension or an ordinal dimension is equally valid. In this case, the decision of which one to use should be based on the characteristics and preferences of the user, so that the selected visualization dimension is as salient as possible for the user. However, if the information to be illustrated in the visualization expresses some kind of order, or can be sorted according to a certain criterion, then the visualization dimension used to map this aspect has to maintain these capabilities, and then using an ordinal dimension is required. For example, if document representations have been clustered according to their publication date, it could be interesting to locate the created clusters in a chronological way —ordering criterion—, so that the representation of the newest documents would be located rightmost than those that have been published in previous years. Analogously, the color saturation of the document representations could be used to illustrate the same information, for example by representing all the documents with the same color, but mapping the age of the documents to the level of saturation of their representations. Then, the oldest
documents would be represented with a poorly saturated color, and the saturation would increase as the documents would be newer.

### 4.5 Information visualization tasks

One of the most cited and used guideline when designing information visualization systems is the mantra proposed by Shneiderman (1996), where he exposes which are the seven functionalities that should be present in any visual information visualization system, regardless its domain, its graphical design and the type of the data illustrated:

- **Overview.** The system has to allow visualizing the whole set of items, even though with a low level of detail.

- **Zoom.** The user has to be able to focus the visualization on a set of items that he/she wants to evaluate with more precision. Focus + Context visualization techniques (Cockburn, Karlson, & Bederson, 2008), whose precursor is the well-known FishEye View (Furnas, 1986), provide a solution to focus the visualization on a set of items that may be of interest for the user, but maintaining the context, and then providing an overview.

- **Filter.** It has to be possible to clear from the visualization those items that the user does not consider interesting. This filtering has to be based on the value of some of the items data —in the proposed document’s model this type of data is called "information elements". The user has to be able to specify these value(s).

- **Details-on-demand.** The user has to be able to explicitly ask the system to provide him/her more details about a specific set of items.

- **Relate.** When dealing with data that are somehow related, the system has to be able to illustrate these relationships so that the user can identify them visually.

- **History.** In order to increase the usability of the system by supporting progressive refinement, error correction and permanent feedback of the system’s state, the system has to be able to keep the list of actions performed by the user. This, for example, allows the user undoing an action whose result is not the one he/she expected.

- **Extract.** A system should be able to save in the system, somehow, which are the set of items that the user has found relevant, so that in future visual information-seeking activities, they can be directly accessed without needing to repeat all the actions that lead to their identification.
The proposed model defines a set of information visualization tasks that are reflected in Figure 4.12. In general terms, these tasks cover most of these high functionalities enunciated by Shneiderman in his mantra. As these functionalities are defined at a general level, they have been used in this work but their names have been adapted to the specific domain being investigated. Thereby, both Overview and Focus are covered in the proposed model by the view manipulation task. In fact, in this task are included all the subtasks that imply modifying how the visualization workspace and its content—in this case, the visual representations—are shown to the user. Then, the Filter functionality is directly mapped into the filtering task, where the use of control widgets indicates which conditions must be met by the documents in order to be visible. The Relate functionality is also directly represented by the clustering task whose aim is to visually represent inter-documents relationships. With respect to Details-on-demand, it matches with the deepening task, as it allows obtaining different levels of detail from a set of documents.

In addition to these tasks, in this work it is proposed to consider three more tasks when designing an information visualization system. First, as many of the data included in the document concept can be sorted, special attention has to be paid to the tasks that allow sorting a set of documents based on a given ordering criterion. Additionally, the qualitative study has confirmed that one of the most repetitive and unsupported task—or supported, but in a non-usable way—while seeking for information is the annotation task. In our opinion, this task has to be considered in the design of the visualization system, as annotating a document does not only means writing a series of notes or summaries, but also assigning tags, indicating that the document is a favorite, or underlining relevant terms, that can be used in the visualization to facilitate the classification, recognition and processing of the documents represented. Finally, a selection task is included as, even if it is more a means than a visualization task—as it is mainly used as a previous required step to perform others tasks like obtaining details-on-demand or annotating a document—, it has a great influence in the visualization as it allows the user to understand with which set of document representation he/she is
working at a given moment. Additionally, in a context where hundreds or thousands of documents have to be represented at the same time, it is not obvious to design one or more selection techniques that allow the user selecting the document(s) he/she wants in an accurate and efficient way.

In Calimaco, for example, are implemented some of these information visualization tasks, like allowing the user to select some documents representations in order to perform other tasks like obtaining details-on-demand, changing their visibility state by defining a set of filtering criteria, classifying them based on the topics they deal with, or changing the position and/or the orientation of the view to focus on a specific part of the visualization workspace. As reflected in Figure 4.13, Google Scholar also provides some of these tasks: getting more information about a result, filtering the obtained results, manipulating the view —by moving to a previous or next page to see ten more results—, and ordering the results according to a given criterion. Finally, 3D Explorer allows clustering the documents according to their content and opening a document by double-clicking over its representation, which can be considered a deepening task. It also allows, as mentioned before, to manipulate the view.

The only functionalities suggested by Shneiderman that have not been considered in the proposed information visualization model are *keeping the history of actions* and *extracting a subset of documents from the visualization*. They have not been included because the approach used in this research is to define only the tasks that have a direct or indirect effect on the visualization. Nonetheless, a visualization system should incorporate both functionalities. In fact, as it will be explained in Chapter 5, each of the actions performed by the user in the system has to be saved, but not only to improve its usability, but also all to be able to model the user, his/her preferences, and his/her actions so that the system can adapt to his/her specific needs at every moment, which is very related with the extraction task. Some mechanisms are also proposed to identify which documents are relevant for the CS researcher according to many aspects, like his/her history of actions, the concrete relevance he/she may provide to documents, or the relevance of some of the entities related with the documents, like its authors or publication venue. More details are provided in Chapter 6. Both cases, nonetheless, can be used in future information visualization activities to adapt the visualization to the user’s preferences.

As it happens in the information-seeking process model explained in Chapter 3, the model of the information visualization system is also flexible and scalable enough to admit the addition of new concepts, like new information visualization tasks, and to modify or delete those already existing without losing its validity and usefulness. As an example, the information visualization tasks implemented in Calimaco to modify
(a) Manipulation of the view in order to visualize earlier or later results; (b) Alphanumeric representation of the documents; (c)-(f) Alphanumeric control widgets used to perform filtering tasks; (d) Document’s related information is displayed using an (other) alphanumeric representation; (e) The visualization dimension “font property” is modified in order to distinguish between the different data related to the document that are illustrated in the workspace; (h) Graphical representations used to illustrate control widgets allowing to perform some tasks.

Figure 4.13 Example of information seeking and visualization system: Google Scholar
Despite not all the possible information visualization tasks that may exist in an information visualization system have been considered, it can be asserted that these tasks mainly can produce three types of effects on the visualization, that are illustrated in Figure 4.15 and explained below.

- A new visual representation is added to the visualization workspace. For instance, a new graphical representation, like a container, can be drawn in the visualization workspace in order to illustrate the result of a clustering task. For example, 3D Explorer allows representing a specific query as a new document whose representation is located in the workspace according to its thematic similarity with the documents. This task, even if is not explicitly considered in the proposed
model, implies the appearance of a new visual representation that represents the query —see Figure 4.2.f.

- Some of the visual representations are somehow affected by the performance of the task:
  
  - One of the visualization dimensions of some visual representations is modified. For instance, the size of the document representations can be modified in order to illustrate the number of pages they have. Something similar is done, for example, both in Calimaco and 3D Explorer, where the color of document representations is used to illustrate the membership to a cluster.
  
  - Sometimes, the same information can be displayed in various ways, and then the visual representations representing it can be illustrated using different modalities. A clear example of it is the representation of the documents and folders in the file explorer of an operating system, like Windows Explorer or Apple Finder. In these environments, which can actually be considered as information visualization systems, the elements can be represented either by icons —that is, visual metaphors— or by textual entries — that is, alphanumerical representation— , depending on the visualization mode selected by the user.
  
  - Modification of the visibility state of a set of visual representations. For instance, the user may want to visualize only those documents he/she has marked as favorites, and then the rest of document representations have to become invisible. As an example, in Calimaco the user is allowed to select and remove from the visualization a single document representation, a set of them or all those that belong to a cluster.

- The visibility state of a visualization workspace can change. As an example, if a deepening task is performed, a new visualization workspace displaying the content of the affected document can appear in the visualization. This occurs in 3D explorer, where the second display can be visible or not depending on whether the user activates it or not.

4.6 Inter-document relationships

As it has been already commented, one of the most useful feedback that can be obtained by visualizing a set of items in an information visualization system consists in identifying how these documents relate. In this case, where items have a lot of data that can be
used to establish relationships between documents, this functionality becomes essential. This is the reason why it has been considered interesting to include in the model which aspects have to be taken into account to illustrate relationships in an information visualization system. First of all, it is essential to know which kind of data is going to be used to determine if two or more documents are related. This criterion is based on the value of a document property —see Figure 4.16— that can be of two types, namely any of the information elements that are included, implicitly or explicitly, in it, like the authors’ names, the topic, the publication date, or the publication venue —see Section 3.3 in Chapter 3, but also any information personally added by the user, like personal notes, tags, highlighted terms, marking a document as favorite, or any other information the user may have added to a document to easily remember its content.

In any case, the relationship is illustrated in the visualization workspace either by the modification of a visualization dimension, or by adding a new visual representation. In the first case, all documents sharing the same value for a given property, are illustrated using the same visualization dimension value. As an example, Calimaco makes use of the color to represent the clustering of the documents. In the second case, the new visual representation is used to visually relate the concerned documents. Getting back to the previous example, the clustering could be represented, for example, by drawing as many containers as clusters, and locating into each of them the representation of the documents of assigned to the corresponding cluster.

Which of them can be used depends on which datum is used to establish the relationship, as some of its features, like its type or the number of values it can take, can impose some restrictions on the inter-documents representations. For example, it has to be determined if the selected dimension has enough distinguishable values to illustrate all the different possible values of the document property that have to be displayed. As an example, if the user wants the system to visualize how a set of documents are related based on their publication date, using the size as visualization dimension would probably not be the best choice as it would imply to have as many distinguishable sizes as publication years, which could be several tens. In this case, for example, it would better to use the spatial position to locate the documents chronologically, or to draw
containers—one per publication year—that would enclose the representation of the documents published in the same year.

However, it is not enough to know which type of information is going to be used to identify if a set of documents are related and to know how the relationship is going to be illustrated. After this, it is needed to specify how each of the values of the document property is going to be mapped into the visualization. In the case of modifying a visualization dimension, it is needed to specify what value of the visualization dimension is going to be assigned to each of the values of the document property. For instance, if color is used to illustrate the topic treated by a document, the visualization system has to assign a specific color to every possible topic. In Calimaco and in 3D Explorer, for example, this assignment is done randomly. On the other side, if a new visual representation is used to illustrate the inter-documents relationship, it is then required to specify what kind of visual representation is going to be added to the visualization. This new visual representation can be predefined by the system and be non-modifiable by the user, in which case the representation does not need any more specification. However, the new visual representation can also have a set of modifiable visualization dimensions that can be used by the system to reflect some kind of information. In this case, the mapping between the document property values and the visualization dimension values has also to be specified. For instance, the system could simply draw a line linking two document representations in order to illustrate that one of them cites the other one. However, it could also modify the lines thickness to illustrate how thematically similar are the two linked documents.

With respect to the assignment of visualization dimension values to document property values, there are two possibilities. First, the user may be asked to explicitly define one by one which visualization dimension value he/she wants to associate to each of the possible document property values. Otherwise, the system has to automatically associate them, either doing it randomly, or making use of the user preferences and of history of use the user has made of the system. The same occurs with the choice of the visualization dimension. The user may be the one in charge of specifying with which visualization dimension he/she wants to illustrate a given document property, or on the contrary the system may automatically define this association. As an example, a user may decide to just indicate that he/she wants the system to use colors to illustrate the venue where the documents have been published, but he/she could also want to further customize the inter-documents relationship representation by indicating that he/she wants it to color in red all the documents that have been published in a given journal X, in green all the documents from journal Y, and let the rest of the documents representations in the predefined color. Calimaco and 3D Explorer are not
flexible enough in this aspect, as they automatically use colors and spatial position as visualization dimensions to illustrate cluster membership and thematic similarity, and do not offer the possibility of indicating how colors or positions have to be assigned to clusters and degrees of similarity, respectively.

4.7 Related works

During the last two decades, information visualization has been extensively studied and the state-of-the-art in the topic has notably evolved. One of the first authors that laid the foundations of the modern information visualization was the French cartographer Jacques Bertin who, based on his experience, identified some of the essential concepts intervening in an information visualization process (Bertin, 1981): marks — that is the graphical objects used to represent the data in the visualization — and its positional, temporal, and retinal properties. The positional property relates to the dimensionality used to illustrate the mark; temporal properties deal with the animation aspects, like mark’s movement; and retinal properties encompass all the visual features used to visually illustrate some information — for example, color, shape, size, saturation, texture and orientation. In the present model, marks are named visual representations, and visual features are called visualization dimensions, but they refer exactly to the same concepts. The only difference remains in the temporal properties, as the only one considered is the movement, which is also considered a visualization dimension. This occurs as, in the proposed model, animation itself is not considered, and movement is then perceived as a form of pattern that may provide some information. However, in the proposed model the movement itself is not associated to any specific meaning, as it only serves to provide some clustering or filtering information, which are the main objectives in information visualization. As an example, applying a movement — like small jumps — to a visual representation could be used to illustrate that the associated document matches a set of filtering criteria, or the direction of the movement could be used to visually cluster the documents — for example, visual representation that vibrate horizontally belong to a cluster, while those vibrating vertically belong to another one.

A few years later, J. Mackinlay (1986) added the enclosure and connection properties as a means of illustrating relationships between objects. The proposed model gives a special weight to the inter-document relationships representation, as both properties are considered in it through the addition of a new visual representation, like a line, that can encompass or link, as appropriate, all the related document representations. MacEachren (1995) and Healey et al. (1996) also expanded the types of visual features that can be used to transmit information, including new elements like color value, color
hue and color saturation, transparency, width, curvature, and flicker, among others. Most of these properties are also included in the proposed model, and those that are not could be easily added.

In the literature, these visual features have been divided into four main groups (Card & Mackinlay, 1997). First of all, there are features related to the use of the space to encode information, like position and orientation. These features have a great importance in information visualization as humans are especially good at processing spatial information. That’s why spatial location is commonly used to illustrate relationships between information, as there is a set of organization principles—called *Gestalt Principles*— that affect the visual perception (Sprenger, Gross, Eggenberger, & Kaufmann, 1997). As an example, two objects that are very close to each other are intuitively considered more similar—or more related—by the observer than two objects that are distant from each other (Blach, Wenzel, Dangelmaier, & Frohmayer, 2007). In the prototypes that have been developed in this work—see Chapter 7—, position is an essential visualization dimension as it reflects the thematic similarity between the documents.

There is a second group of some properties that are called retinal, as the retina of the eye is directly sensitive to them, regardless the rest of the visualization. These properties are used either to illustrate some information of the data, at the individual level, or to reflect how these data are related. In this group are included visual features like the color or the size.

In this sense, there is a third type of features that allow illustrating relationships between data: connection—two elements that are connected by a line, for example, are clearly perceived as related by the observer—, and enclosure—a set of elements that are located in the same zone which is, explicitly or implicitly, delimited by borders, are also perceived as related by the user.

Finally, in case that animation is used in the visualization, the variation of the visual representation along time also allows encoding information about the data. As an example, the rotation speed of a visual representation can provide some information to the observer.

All these visual have a main characteristic in common: they are preattentive features. As mentioned before, this aspect is very powerful, as it ensures a fast and efficient impact on the user, but without cognitively overloading him/her. In the proposed model, most of these visual features have been contemplated, but visualizing and organizing them in a different way, as instead of categorizing them based on their nature or on the cognitive ability that the user uses to identify it, they are classified according to the type of data they can illustrate—ordinal or nominal. Moreover, in the proposed model, the relation
of these features with the rest of concepts involved in the information visualization is clearly indicated. Then, even if all approaches are correct and complementary, we believe that the proposed model is the most simple and straightforward approach in order to design an information visualization system, as it does not take into account the physiological or psychological aspects of the visual properties, but only their possibilities and limitations of use.

Even if the basis is the same for almost any information visualization system, the way in which these marks and its visual features are used to transmit the information is very variant. In fact, many authors have stated that visualizing information is a domain-specific activity as there are many aspects that have to be considered when designing an information visualization systems (Jaeschke, Leissler, & Hemmje, 2005; Jaeschke, Gupta, & Hemmje, 2005; Wiss, Carr, & Jonsson, 1998), like the type and structure of the data to be displayed (Siirtola, Räihä, Säily, & Nevalainen, 2010; Wu, 2010; Sugibuchi, Spyritos, & Siminenko, 2009; Ren, Tian, Zhang, & Zhang, 2010; Semiu & Jamaludin, 2013), the user—in terms of perceptual and cognitive capabilities, but also of preferences and needs— (Amar & Stasko, 2004, 2005; Carr, 1999; Liu & Stasko, 2010; Nowell, Schulman, & Hix, 2002), and the context (Kirste & Rauschenbach, 1996; Aaltonen & Lehikoinen, 2005; Bugajska, 2005; Craig, Huang, Chen, Wang, & Zhang, 2015). In this research, it has been considered that many aspects are pretty much common regardless the domain—for example which visualization dimension can be used to transmit a given information—, but there are others aspects that are inevitably domain-specific, like the representation of the relationships. In fact, the criteria used to relate two elements depend first of all on their nature. It is not the same to illustrate the similarity between two complex entities, like documents, as they have many facets—authors, topic, relevance, size...—, than relating two simple numerical values. In the first case, it has to be decided which facets are taken into account to define the similarity, and if there are many how their similarities are illustrated so that the user can recognize and understand them intuitively. Moreover, some mechanisms have to be used to calculate the similarity between two entities. In the case of numerical values, on the contrary, the comparison between two elements is direct, universal and unique, as a given number is always, under any circumstance, less than, equal to, or greater than another number. Then, it may be considered that visualization does not only depend on the domain, but also on the nature of the data to be visualized and of their relationships. That’s why the present model, even if it can be very useful in many domains, has some specificities related to the scope of the study—in this case, the information-seeking process in the research domain. Finally, in this work we also admit that it is essential to take into consideration the user’s capabilities and preferences.
in order to properly model and design an information visualization system. For that reason, in the researcher’s model presented in Chapter 5, preferences related to the visualization have a special weight.

But, regardless the domain in which an information visualization system is used, the data that are visualized or the relationships that can be represented, or the user who is visualizing the information, there is some consensus about which is the sequence of steps that is required to be done in order to visualize information. The reference model for visualization of Chi and Riedl (1998) is one the most recognized and used in the field. This model, as reflected in Figure 4.17 consists of three main stages and four basic concepts. Mazza (2009) and Card et al. (1999) propose similar models with, in general terms, the same stages and elements.

First of all, the unprocessed raw data is translated into a computationally understandable representation, called data tables. At this point, it is required to define which visual structures—that is which marks and graphical properties—are going to be used to visually represent the data contained in the tables. Finally, how the user is going to be able to visualize the environment and its content is defined in the last stage, where user decides how to manipulate the viewpoint in order to carry out, for example, a focus + context view (Furnas, 1986) —the view focuses on a specific information, but the context of it is still somehow displayed—or a overview + detail view (Shneiderman, 1996)—there are two visualization environments, one of them displaying an overview of the whole visualization workspace, while the other provide more in-detail information about a concrete visual structure. Even if the proposed model is not sequence-oriented as the this reference model, all the concepts mentioned in the later appear in our solution, as it can be seen in Figure 4.18.
Later on, Aaltonen and Lehikoinen (2005) adapted this reference model to make it capable of taking into account the context in which the visualization activity is carried out. To do this, they propose to incorporate a set of simple computational-like rules in the data transformation stage so that, depending on the context, raw data is processed one way or another. Additionally, they propose to create several several layers of visual structures, so that the visualization varies according to the context.

In both cases, the models present some of the relevant concepts intervening in an information visualization activity, but their main purpose is to represent the sequence of steps that have to be done to illustrate a set of data. The purpose of this research, on the contrary, is to determine which are the relevant concepts that intervene in the process, and how they relate. However, both types of models complement each other, as they provide a different point of view of the information-seeking process. In fact, some authors have followed a similar approach, like Bugajska (2003, 2005) who provides a complete framework of the information visualization process that contains many concepts intervening in it —the user, the tasks, the visual representations and their visual characteristics, the navigation of the visualization workspace, the environments in which the visualization is performed, etc.— and their relationships —between visual representations, between the user and the visual representations, between the context and the visual representations, etc. Jaeschke et al. (2005, 2005) also identify several relevant aspects related to information visualization and provide a standard and specific language to formally make use of them in the context of information visualization.

With respect to the tasks that have been included in the model, it has been already explained that they cover the basic functionalities that should be considered in any information visualization system based on the Shneiderman’s mantra (Shneiderman,
At a more detailed level, the model also takes into account some more specific tasks that are related in some way to the visualization, like sorting the items, selecting them, manipulate the view—for example allowing the user to navigate within the digital environment—, providing multiple visualization workspaces, or annotating something in the visualization workspace (Heer & Shneiderman, 2012; Brehmer & Munzner, 2013).

Finally, to the best of our knowledge, nobody has provided a conceptual model illustrated through a visual, standard-based, and non-ambiguous representation, containing all the main concepts and relationship that are somehow implied in an information visualization activity, not in terms of temporal sequence, but in terms of semantics. Additionally, as it occurs in the information-seeking process model presented in Chapter 3, using UML to represent the model grants it with a great flexibility, as it allows to add, modify and delete concepts and relationships to adapt to new realities, or to include new aspects that may have been unconsidered.
Chapter 5

Model of Computer Science researchers

Undoubtedly, the component that introduces most complexity in the described information-seeking process is the user, which in this case is a CS researcher, because of the difficulty associated with trying to model human characteristics and behaviors. As mentioned in Chapter 3, several authors have investigated the information-seeking process from the user’s perspective, usually in terms of emotions, feelings and thoughts, like anxiety, uncertainty, doubt, pessimism or discouragement at the beginning, and satisfaction, motivation, pride, relief, optimism or completeness in the last stages of the process (Kuhlthau, 1991; Wilson, 1997; Ford, 2015).

This research does not focus in the psychological aspects of the CS researcher, nor in their evolution along the process. In this thesis, however, a conceptual model of the CS researcher is proposed based on a set of characteristics that are directly related to the performance of the tasks leading to the achievement of an information-seeking activity. In fact, it is essential to properly model the CS researcher, as the information-seeking process clearly depends on him/her. Ford (2015) clearly states that there is a set of internal and external aspects related to the user that directly affect the performance of information-seeking activities, like demographic, cognitive, and affective aspects in the first case, and features related to the work, education, leisure, social relationships and user’s role as a citizen. Many authors have also taken into account some of these aspects to model the information-seeking and information behavior process (Kuhlthau, 1991; Dervin, 1992; Savolainen, 1995; Byström & Järvelin, 1995; Wilson, 1997; Sutcliffe & Ennis, 1998; Case, 2012).

One of the best examples of the need of properly modeling the user is the field of adaptive hypermedia. Brusilovsky (1996, p.2) defines adaptive hypermedia systems as "all the hypertext and hypermedia systems which reflect some features of
the user in the user model and apply this model to adapt various visible aspects of the system to the user". This definition perfectly match the ultimate goal of this research, which is design and developing a system able to adapt its behavior, interaction and visualization to any aspects of those that intervene in the information-seeking process, like the user. Then, according to this definition, it could be considered that the final goal is to develop an adaptive hypermedia information system. In fact, Brusilovsky (2001) defines information retrieval as one of the fields where adaptive hypermedia systems can be more suitable, either by providing adaptive search engines, adaptive guidance systems, adaptive recommendation systems, or adaptive bookmark systems.

Brusilovsky (1996, 2001) states that there are seven main components that have to be included in a user profile in order to make it useful to design adaptive information systems:

1. User’s goals. This component matches with the "information-seeking purpose" proposed in the high level information-seeking process model explained in Section 3.2 of Chapter 3;

2. User’s knowledge. This component has been included in the CS researcher’s model presented in this chapter as the concept "Relevant knowledge" —see Section 5.1.1;

3. User’s background and experience. This component is also been present in the CS researcher’s model presented in this chapter through the concept "Information-Seeking Heuristic" that derives from the CS researcher’s "Relevant Knowledge Related to the Information-Seeking Activities themselves" —see Section 5.1.1;

4. User’s preferences. The CS researcher model explained in this chapter also consider the user’s "Personal Preferences" as a component that has to be known to properly describe the user —see Section 5.1.3;

5. User’s interests. This aspect perfectly matches with the thematic interest of the CS researcher described in Section 5.2 of this chapter;

6. User’s individual traits. "Personal Characteristics", both physical and psychological, have also been included as a component of the CS researcher’s model —see Section 5.1.2;

7. Context-related aspects. In this category are included aspects like user’s location, or the platform used by the user —both hardware, software. In this research, these aspects have also been considered, but they in the context’s model —see Section 3.6;
Although the components of the user’s model have been described quite clearly in the general field of adaptive systems, to the best of our knowledge, there is no specific research to model the user in a specific domain, that is research-oriented information-seeking, and in a specific field like computer science. In order to fill this gap, in this chapter are illustrated all the characteristics of these specific domain- and field-dependent users through a UML-based conceptual model, as done for the information-seeking process —see Chapter 3— and the information visualization activities —see Chapter 4. Appendix C.5 illustrates the full model of the CS researcher.

5.1 Main components of the researcher’s model

Based on the results of the qualitative study, four aspects have been considered relevant in order to model a CS researcher —see Figure 5.1. First of all, each individual has a demographic profile including personal information like his/her gender, his/her nationality, his/her age or his/her 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 —from where he/she has acquired some relevant knowledge—, and has a set of specific personal preferences based on his/her tastes. Finally, each researcher, as a human individual, has a set of personal characteristics related to his/her way of being, capabilities and restrictions. Most of these aspects have been identified by other authors as factors that affect the information seeking process. In concrete, Ford (2015) defines a set external and internal factors that influence the process. The first ones are related to the context in which the seeking is carried out, while the others have to do with the demographic, cognitive and affective aspects of the CS researcher himself/herself. In the proposed solution, however, as adaptation and personalization are very relevant features, a special importance is given to the researcher’s preferences. The same occurs with the knowledge that the CS researcher has, as seeking information in a research context is a very focused process that is highly dependent on the topic and on the level of knowledge the CS researcher has in it. This second aspect is further developed in Chapter 6. Both aspects are defined by Ford as cognitive aspects.

5.1.1 Relevant knowledge

One of the CS researcher’s aspects that most influences the information-seeking activities is the amount of relevant knowledge he/she has. As illustrated in Figure 5.2, part of this knowledge is related to the information-seeking activities themselves, while the other part is related to one or more research topics, as a CS researcher can already have some knowledge about the investigated topic(s), like who are the reference authors,
which are the reference research groups that do research on it, which are the reference documents and/or publication venues dealing with it, or what its specific terminology is.

Even if this is a rough simplification of a much more complex construct that goes beyond the scope of this research, the analysis of the qualitative study has allowed to identify two main researcher’s profiles that are partly based on the amount of knowledge—both procedural and thematic— he/she has acquired, as reflected in Figure 5.3.

A junior CS researcher usually lacks—or has very little—previous experience in seeking and finding information in a research context, resulting in lower levels of self-confidence and intuition.
A senior researcher, on the contrary, is supposed to pay more attention to his/her intuition as he/she has acquired a lot of knowledge related to the information-seeking activities themselves stemming from his/her experience in looking and finding information. Moreover, a senior researcher usually relies a lot more on himself/herself due to both his/her experience and the knowledge about a topic he/she has accumulated throughout his/her career.

With respect to the procedural knowledge, the interviewed CS senior researchers were able to more clearly explain which steps they usually carry out when seeking for information, as they follow a more standardized process based on their experience, whereas junior researchers were more hesitant when expressing how they perform the seeking, and especially explaining why they take one or another decision. In practice, this procedural knowledge translates in the acquisition and use of some heuristics 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, but, even when they realized they actually use some of them, and after thinking about it for a while, 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 then are difficult to be modeled. 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.

The difference in the level of acquired procedural knowledge also has a consequence on the number of tasks performed by the CS researcher, as junior CS researchers often perform a lot more information-seeking tasks than senior CS researchers do, as the first ones lack self-confidence and then they almost systematically doubt about the accuracy and relevance of the results they obtain, and then repeat the process many times to confirm them. Additionally, junior CS researchers are often more likely to fear missing some relevant information, which makes them perform very exhaustive searches to ensure the completeness of the results. Senior CS researchers, on the contrary, are faster determining the relevance and usefulness of a document, and are less fearful of leaving some relevant document untreated, as, if the document is really relevant, then they probably already know it or at least some of its information elements —like its authors. This, for example, occurs with the example of the qualitative study explained above, where the senior CS researcher does not fear considering only the first 15-20 results of a filtering task.
On the other hand, the thematic 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 during an exploration task. 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 scientific 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 activity. In this regard, it is important to distinguish expertise and relevance, as a CS researcher can be an expert in a topic, but yet it may not be a reference author in it because, for example, he/she may have not produced any relevant contribution to the scientific community.

5.1.2 Personal characteristics

Among the personal characteristics of a CS researcher that have to be taken into account to model the CS researcher, there are physical and psychological aspects —see Figure 5.4.

![Figure 5.4 Personal characteristics of a CS researcher](image)

The first group refers to all the CS researcher’s physical aspects that may affect, at some level, the performance of the information-seeking activity. Considering these aspects is essential not only to ensure the accessibility of the system for users with functional diversity, but also to allow the system adapting to the specificities of the CS researcher and then improving his/her user experience while using the system. This is especially relevant in highly interactive systems, as visualization and interaction are core elements in the process, and then all the user’s characteristics related to vision and motion have to be considered in the CS researcher’s model. In the first case, for example, the system has to make good use of colors to facilitate, in general terms, their recognition and distinction, but also to take into consideration that some users can have
some form of color blindness when choosing the palette to be used and defining the goal of each color. Similarly, an information visualization system has to allow reducing the focus of attention so that users who have some type of peripheral vision problem can also make use of it in a comfortable and usable way. This does not mean that the system has to limit the use of visual characteristics in the system, but that visualization has to adapt to user’s needs and preferences. With respect to interaction, the system has also to consider that there are users with different physical abilities and limitations, and then it has to be capable to adapt to their specificities. This means that, for example, a system should provide different ways to perform the same action, but with different levels of interaction complexity, ranging from simpler—even if slightly less efficient—to more complex interaction—and probably more efficient. By doing this, the system takes into consideration both the different types of users in terms of experience, and the restrictions of those users with some kind of physical functional diversity, as for example having a low level of hands dexterity.

The second category of personal characteristics groups all the psychological aspects of the CS researcher. Within this category, it can be distinguished between personality traits and cognitive characteristics. Even if the present work does not focus on the information behavior of CS researchers, some concepts have aroused as a result of the qualitative study, and then they are briefly presented. However, 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 from a more 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, as it will be exposed later—, only the concepts mentioned by the interviewees are described, letting a more in-depth analysis for future works.

Personality traits make reference to the inner behavioral aspects that are related to the CS researcher’s personality, like how much self-confidence he/she has both in general and while performing an information-seeking activity. This trait is in line with one of the affective factors defined by Ford (2015), self-efficacy, that allows him to conclude that the “individual’s belief in his or her ability to successfully perform some task” is a relevant aspect that can condition the fulfillment of an information need. In fact, Ford states that personality is itself a relevant factor in information behavior that affects essential aspects of the process, like the strategy followed by the CS researcher to achieve his/her information-seeking purpose. In this research, the analysis of the qualitative study reflects that intuition, together with self-confidence, are two of the personality traits that define the CS researcher’s behavior when performing an information-seeking activity. In fact, even if many of the interviewees reported that
they have gained self-confidence and have resorted more to their intuition —almost unintentionally— as they have undergone information-seeking activities once and again, their comments also reflected that their own personality greatly nuances the effects of experience.

Apart from this, there are several cognitive personal characteristics that are related to the processing and understanding of information leading to the construction of new knowledge in a learning process like information-seeking. Regardless the learning process, each individual assimilates, associates and constructs new information in different ways, called cognitive or learning models. As seeking for information in general, and in research in concrete, is mainly a sense-making and knowledge-building process (Dervin, 1983; Dervin & Nilan, 1986; Dervin, 1992), the mental model followed by the CS researcher is a relevant factor that has to be taken into account when modeling him/her during the information-seeking process (Ford, 2015). Recognizing this mental model and adapting the system to it is an essential step to ensure higher levels of learnability and ease-of-use, leading to the increase of the system’s usability. For instance, the type of learning model directly affects how much help or suggestions does the CS researcher want to obtain from the system, and at what moment. As an example, while some users may feel comfortable with the system providing them with some help to perform a given task, others may prefer to do it just by themselves.

Beyond this, there are other cognitive aspects that are part of the CS researcher’s model —especially when visualization and interaction are so relevant in the process—, like the level of field-dependence of the user when observing a great amount of interleaved visual elements, how many information can be provided simultaneously by the system before overloading the CS researcher in terms of cognition, or how complex can the visualization be before becoming confusing. This, for example, clearly affects how many visualization dimensions can be simultaneously used at maximum to ensure a good level of usability and a great user experience.

5.1.3 Personal preferences

Each CS researcher is unique, and then has different personal preferences that fit with his/her particular and individual point of view when it comes to perform an information-seeking activity. 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 specific and isolated preferences that may be different from the previous ones, as they are influenced by the specific information-seeking purpose he/she wants to achieve, by the information-seeking tasks he/she performs, and/or by the context in which the activity
takes place. The personal preferences that the participants of the qualitative study have directly or indirectly mentioned are illustrated in Figure 5.5 and are explained below. They are probably a small sub-set of all the possible personal preferences that may exist, but, as already mentioned before, the proposed model is flexible enough to accept new preferences without affecting at all the rest of the model.

![Figure 5.5 Personal preferences of a CS researcher](image)

In order to design an adaptive information visualization system that takes into account these CS researcher’s characteristics and preferences, but also his/her information-seeking purpose and the context in which the information-seeking activity is performed, it is required to define which data are used as input, when do these data have to trigger a modification in the visualization, and which kind of modification has to be performed.

An information visualization system can obtain these personal preferences by two means. On the one hand, the CS researcher himself/herself can provide some specific and direct information to indicate which are his/her preferences. This approach ensures that the system properly models the user in terms of preferences, but it also increases the his/her workload, as it obliges him/her to perform system tasks instead of user tasks. Even if this approach has to be avoided as long as possible in order to increase the system’s usability, it may be difficult for the system to infer some of these preferences, and then it may be better to ask the user to provide some information —like some personality traits, or his/her main priority when seeking information.

The second approach consists in tracking how the user interacts with the information system, in order to later infer his/her preferences by processing the collected data. In most of the cases, these preferences can be identified just by creating a set of counters that are increased when a given action is performed by the user under a certain set of conditions. As an example, a counter can be defined for each of the document’s information element in order to increase them when the corresponding information element is used by the CS researcher to define a filtering mechanism. Then, at every moment, the system can know which are the information elements that are most frequently used by the CS researcher when filtering a document collection. Similarly, this can be done with most of the preferences.
The following sub-sections present some of the preferences, together with some of the data that should be registered by the system, transparently to the user, in order to infer them. In Chapter 6 is explained a concrete use of the data, consisting in estimating the relevance of documents, authors, publication venues and research groups based on the CS researcher’s preferences and interests. Finally, in Chapter 7 are presented some examples of how the visualization can adapt to the value of these data.

5.1.3.1 General preferences

Beyond the needs that are implicitly associated to each information-seeking purpose, a CS researcher may have some personal preference related to the information-seeking activity in general —see Figure 5.6.

![Figure 5.6 CS researcher’s preferences related to the information-seeking activity as a whole](image)

To be more specific, 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 prefer to obtain exhaustive results. In the first case, the CS researcher prefers to get and process all the documents that can be useful at some level to achieve his/her information-seeking purpose, even if this may imply analyzing and even keeping some documents that are actually not relevant for the current purpose. In the second case, the CS research usually prefers obtaining a smaller subset of results provided that they 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, senior CS researchers usually prefer being precise as their searches are more directed, they are more confident taking quick decisions, and they usually lack time, while junior CS researchers normally perform more exploratory searches and lack relevant knowledge about the topic(s) they investigate, which makes them feel more comfortable if they are exhaustive in the information-seeking process.

Moreover, along the information-seeking activity, each CS researcher prefers to gather different types of information during the process, depending on how he/she
prefers to address the problem. Then, for example, he/she may prefer to take notes about:

- the terms that seem to be relevant in the topic in order to use them to refine the query in the next chained filtering task;
- some information element values that, according to him/her, may be relevant to successfully explore a document collection;
- or the terms and information element values that actually fulfill their information need.

The last general aspect that has to be considered when modeling the CS researcher is the language he/she prefers to use both to perform the information-seeking process—for example the language selected for the search engines, or for the information visualization system used to seek information—, and to read the documents. Usually, CS researchers only consider documents written in English even if it is not their mother language, as it is the international language that is mostly used in research, and then it is usually the one used in order to reach as many people as possible. However, some of the interviewees also expressed that, occasionally, they also consider documents in other languages, if they believe they can be relevant, as long as they are able to understand them or to translate them into a known language.

### 5.1.3.2 Context-related preferences

As explained in Chapter 3, the context has to do with all the external aspects that interfere, directly or indirectly, in the performance of the information-seeking activity, like the time and the space in which it takes place. Researchers can also have some preferences related to these aspects, as reflected in Figure 5.7.

![Personal Preference related to the Context](image)

**Figure 5.7** CS researcher’s preferences related to the context in which are carried out the information-seeking activities

First of all, every CS researcher configures his/her workplace according to his/she needs, but mainly to his/her preferences. Some CS researchers make use of two or more
computer monitors as they like to enlarge their visualization and interaction workspace, while others only want to use one. There are even some intervieweees that expressed that, occasionally, they use other devices, like smartphones or tablets, to perform part of the information-seeking activity —usually to read, but sometimes even to filter or explore document collections. In fact, this is becoming more and more common as content is getting fully, easily and efficiently accessible from mobile platforms due to the implementations of responsive webs and applications where the visualization of the content adapts to the device. Besides, many editorials —like ACM or IEEE— have implemented mobile applications allowing to access and read their papers, and many of the digital tools that can be used by a CS researcher, like reference managers, have also launched applications that provide their services in a usable way —for example Mendeley. This makes increase the amount and type of combinations of workspaces that can be used, which together with the CS researcher inner uncertainty, conforms the need of capturing which are the preferences of each CS researcher in this aspect. As an example, a CS researcher can use more than one computer screen and use them as a single expanded monitor, whereas others can use each monitor for a different purpose —for example one for reading, and the other one for filtering and exploring.

Apart from this, each workspace can be formed by many work areas. A work area can be defined as a physical or digital space where the CS researcher performs a specific kind of action —like reading, filtering, exploring, annotating, writing, and so on. In summary, a work area can be seen as a subdivision of the workspace. As an example, many of the intervieweees explained that, while seeking information in the computer, they take notes in a digital document. In this case, the visualization workspace and the digital document can be considered different work areas that are used for different purposes.

On the other hand, tools used to perform the information-seeking activity —or at least part of it— are also part of the context, and then which of them are preferably used by the CS researcher also reflects his/her preferences. For example, some CS researchers may prefer to start the seeking by using a meta-search engine like Google Scholar, while others may prefer to be more concrete and use the search engine of a specific digital library.

A CS researcher may also have certain preferences related to temporal aspects, like if he/she prefers to carry out the information-seeking activities at a given moment of the day —some, for example, may prefer to do it early in the morning—, how much time he/she uses to spend seeking information, or if he/she prefers carrying out the information-seeking activities without stopping, or splitting it in various sub-activities dispersed in time.
Finally, it can be useful to know if the CS researcher commonly looks for information alone, or if the information-seeking activity is usually distributed between several researchers. In this case, the system should record who are the other researchers, in which way they collaborate, and which is their preferred communication method.

### 5.1.3.3 Visualization-related preferences

In case the CS researcher makes use of a digital means of access, either a complete information visualization system, or a search engine, the system in question should be able to recognize which are his/her preferences in terms of visualization. This aspect is essential as, even if the system is very efficient and precise in retrieving information, if the information is not visualized in a clear, simple and intuitive way, the CS researcher may feel that the system is complex and confusing, and then he/she may decide not to use it.

As reported in Chapter 4, there are many aspects that have to be considered when visualizing information. Some of them are presented in Figure 5.8.

![Figure 5.8 CS researcher’s preferences related to the visualization of the information](image)

One of the most important is the election of the visual representations that the system is able to use to visually represent the data. In order to adapt to the user and make the visualization more appealing and understandable, the system should know which type of representations the CS researcher commonly uses.

The same occurs with the visualization dimensions used to represent classification, order or clustering. In this case, a CS researched may prefer to use some dimensions over others, as he/she considers them more salient. Besides this, a CS researcher may have the habit of generally using the same visualization dimension to cluster or sort the data according to a given criterion. For example, a CS researcher may like to illustrate the impact metric of documents through their size, while others may use this dimension to reflect their number of pages.

Finally, the load level of visual information from which a CS researcher feels overwhelmed varies from one person to another. Then, the CS researcher’s model
has to contain, for example, how many different visualization dimension he/she uses at maximum so that the system knows where is his/her limit. The same applies to the magnitude of documents he/she usually likes to explore at the same time to feel comfortable. This preference is explained in more detail a little bit later, in Sub-Section 5.1.3.4.2.

Nonetheless, documents are not the only information that can be displayed in an information visualization system. As reflected in the information visualization model, it is likely to display a legend explaining how the visualization dimensions are being used, and then it is worth knowing how the CS researcher makes use of it. This implies, for example, identifying if he/she always use it, or if on the contrary it only makes it visible occasionally. in the latter case, it is important recording in which situation he/she makes use of it. If the system offers different ways to provide the legend, it also can be useful to record which ones he/she prefers.

Finally, and at a more general level, a system can visualize the information through environments with different dimensionality, and a CS researcher may feel more comfortable using one of them over the others. It is also important to identify in which cases he/she prefers using one or another type of environment. As an example, a CS researcher may prefer to use a one-dimensional environment —that is, a list— to display a filtering result, but he may prefer to explore a document collection using a three-dimensional environment. Knowing this preference is essential, as the rest of the visualization depends on it.

5.1.3.4 Preferences related to information-seeking tasks

Finally, a CS may have several preferences related to the performance of the information-seeking tasks, as reflected in Figure 5.9.

In the following sub-sections are presented some of these preferences.

5.1.3.4.1 Filtering preferences

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 some criteria that are needed to achieve the current purpose, and/or by discarding those ones that match the criteria that are considered irrelevant. Which of the two strategies is a priori adopted depends on the CS researcher’s preferences. According to the results of the qualitative study, CS researchers usually follow one or another filtering strategy according to their expertise, and above all to their self-confidence. As mentioned before, junior CS researcher usually lack self-confidence and intuition, and then it is a lot more difficult for them to ensure that a document is not relevant for the current purpose than for a senior
CS researcher, as the first one has the tendency to feel that everything is important and then fears discarding something that might be relevant for his/her purpose.

On the other side, as mentioned before, during the filtering task the selection or discarding of a document relies on the terms selected to define the query string, or on the value of one or more information elements of the document. Again in this case, depending on his/her background, the CS researcher can opt for using one of the two approaches, or even to combine them. In general, the use of information elements in a filtering task reflects that the CS researcher has some relevant knowledge allowing him/her to narrow the scope of the search. As an example, if a CS researcher is an expert on a topic, looking for relevant documents in order to elaborate the state-of-the-art probably would start by delimiting the sources of information to a set of journals that he/she knows that are relevant in the topic. Which information elements are used to filter a document collection depends not only on the knowledge the CS researcher has, but also on the importance he/she assigns to each of the information elements. As
an example, one of the senior CS researcher interviewed affirmed that she usually can
determine if a document is potentially relevant for the current purpose just by reading
its title and authors, as her investigation is very focused and she perfectly knows which
are the specific terminology that is used in it, and who are the authors that investigate
and are relevant in it. Many other interviewees, nonetheless, 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 motivation
that makes a CS researcher continuing with an information-seeking activity. According
to the interviewees, one of these motivations is related to the size of the document
collection being manipulated. In particular, a CS researcher may want to perform as
many chained filtering tasks as needed to reduce the number of resulting documents
until he/she considers the resulting collection is small enough to be explored more in
detail. Then, the magnitude of the filtering result can be considered as a requisite to
stop filtering a collection. This criterion depends on the information-seeking purpose to
be achieved, but also on the CS researchers’ preferences, as the amount of document
that a user considers manageable greatly varies according the user. As an example, a
given CS researcher can feel comfortable with a set of a hundred documents returned
after filtering a collection, while another one can feel overwhelmed by such amount of
documents and may prefer to carry out more chained filtering tasks until he/she gets
as a result twenty documents at most. In the same way, the CS researcher’s model also
includes which are the preferences of use of the CS researcher once he/she has obtained
partial or final filtering results —like for example saving them in a local collection,
reading them during the filtering task itself, tagging them with some labels, marking
some of them as favorites, or printing them. By modeling this, the system has enough
data to be able to infer the possible next steps of the user after filtering the collection,
which would allow to adapt the visualization and the interaction to his/her future needs.

5.1.3.4.2 Exploration preferences

When exploring a document collection, even if a lot of documents are displayed at a
time, a CS researcher typically only turns his/her attention to a small subset of them at
every moment in order to avoid being overwhelmed by too much information. Knowing
this magnitude can be very helpful as the system can vary how much information or
how many documents at maximum have to be simultaneously displayed in order to
adapt to the CS researcher’s preferences.

Apart from it, exploration is conceived as a browsing task, where the CS researcher
makes use of his/her visual capacities to figure out what a collection is about. Then,
it is essential to know which are the user’s preferred visualization dimensions, so that the system can make a preferential use of them. This results in an easier and faster understanding of the visualization by the CS researcher, and then in a more effective and efficient exploration.

Visualization dimensions are used to visually illustrate some of the document’s information, and then they are mapped, based on the CS researcher’s indications, to each of the information elements that have to be visualized. This mapping, then, also reflects some of the CS researcher’s preferences, like which are the information elements he/she usually uses when exploring a document collection, which visualization does he/she prefer to use to illustrate the values of a given information element, or how many information elements, at most, does he/she likes to illustrate at the same through visualization dimensions. As an example, a CS researcher may prefer to examine a set of documents by using colors, spatial positions and sizes to cluster them, respectively, according to their main author, their publication date and the number of time they have been cited, while another CS researcher may prefer to illustrate the documents whose main author is the same by drawing their representations with the same pattern, varying their saturation based on their publication date and changing their size according to the number of citations. Also, another CS researcher could prefer to use at every moment only one visualization dimension, and then he/she should perform the activity in three steps: first filtering the collection according to the colors used to illustrate documents with the same main author; then, the color of the representations resulting from the previous filtering task could be used to indicate the publication date; and finally, the color would be used again, but in this case to illustrate the number of citations. In this second case, it could be inferred that the CS researcher does not like to use more than one visualization dimension, and that he/she likes to map the same visualization dimension to different information elements.

As explained in the previous section, the CS researcher’s model can be greatly enriched by reflecting which are the actions he/she prefers to do when or after exploring a document collection, like recording how many notes are taken during the exploration, identifying which kind of notes does the CS researcher takes, determining if some of these notes are used to define the filtering mechanism of future filtering tasks, tracking if he/she reads, stores, tags, or prints any document during the exploration.

5.1.3.4.3 Reading preferences
Most of the CS researchers interviewed in the qualitative study indicated that, at some point, they read the whole document to determine if it actually is useful for his/her information-seeking purpose and to extract all the relevant information provided by the
document. However, in some cases the CS researcher only needs to read a small part of it as his/her information need is very concrete and he/she does not need to read the full document. In these cases, the CS researcher can have some preferences and may give priority to some parts of the document —like sections or chapters—, as he/she considers they probably contain enough information to achieve their purpose.

Besides, some CS researcher stated that they prefer to read a document in paper format, while others opt for the digital format. Roughly, many of the interviewees have expressed that they perform all the process in a digital context as there are many digital tools that facilitate the access and manipulation of big document collections but, when they have to read and work on a document, they prefer to print it.

While reading, it is common to highlight some parts, or to take notes about the content. Knowing if the CS researcher does it, and if so, what kind of notes does he/she takes or what kind of content does he/she highlights can enrich the model as it allows determining if the CS researcher makes an active or a passive reading. This can help representing on which elements of the documents is the CS researcher specially interested —for example terms, diagrams, tables, or full sections.

Finally, as in the other information-seeking tasks, the CS researcher model has to include information about how the information-seeking activity is usually continued after reading a document.

5.1.3.4.4 Storage & Management preferences
A CS researcher can have some personal preferences related to the storage & management of the documents. First of all, a CS researcher can follow a classification strategy to organize the documents he/she archives. This strategy is partly defined by the set of criteria that are used by the CS researcher to classify, sort or cluster a set of documents. These criteria are commonly based on one or more of the implicit or explicit data of the documents. As an example, one of the interviewed CS researchers explained that he groups his documents according to the publication venue where they have been published, while some others expressed that they classify their documents according to the topic they address, or to one of their authors —either the principal one, or the one they know, regardless his/her authorship position.

On the other hand, when it comes to organize the stored documents, each CS researcher also prefers to structure his/her local collection following a given structure. Based on the qualitative study, four main strategies have been identified. A depth structure is the one where the CS researcher uses a vertical hierarchy with very specific classification levels, even if this implies having many of them containing few documents. In a breadth structure, on the contrary, the CS researcher uses a mainly horizontal
hierarchy with a few levels that are more generic, even if this implies having a lot of documents at each level. In the first case, it can be considered that the organization, being more explicit, makes subsequent search processes more independent on the knowledge of the CS researcher, as each level provides enough information —through the name of the class— to help him/her navigating to the target class containing the documents he/she wants to locate. In the second case, the search relies a lot more on the CS researcher’s knowledge as the organization is more flat and abstract, and then the variety of documents archived in the same class is higher. Figure 5.10 illustrates how a set of documents written by the author $a_1$ would be stored depending on the structural preference used. In the case of the depth structure, within the folder of the author there would be as many folders as topics addressed by the author —$t_1$ and $t_2$— that in turn would contained a sub-folder for each of their more specific sub-topics —$t_{1,1}$ and $t_{1,2}$ in $t_1$, and $t_{2,1}$ and $t_{2,2}$ in $t_2$. Finally, each of these sub-topic folders would contain the corresponding documents: $d_1$ and $d_2$ in $t_{1,1}$; $d_3$ and $d_4$ in $t_{1,2}$; $d_5$ and $d_6$ in $t_{2,1}$; and $d_7$ and $d_8$ in $t_{2,2}$. On the contrary, a CS researcher using a breadth structure would store all the documents in a single folder assigned to the author.

The third classification strategy consists in not classifying, as CS researchers can even not want to store the documents in a structured way, which means there is no structure at all, and all documents are stored together. This often happens with the documents considered potentially useful in the future, as they are all stored in the same folder, with a generic name like “papers to read”.

Finally, CS researchers can base the organization of their local document collections on the assignment of tags. In this case, the organization does not rely on the location where the document is stored, but on the tags assigned to it. Even if tags do not explicitly classify the documents in levels as in classical hierarchies, it can be considered that all the documents tagged with a common label —or a set of labels— virtually belong to the same class. Besides, as every document can have more than one tag.
assigned, the flexibility of the classification increases a lot because the same document can be virtually located in many classes without having to duplicate it.

The second aspect that has to be considered when modeling the CS researcher’s preferences related to the storage and management of documents is the personal storage means preferred by the CS researcher to archive his/her documents. In general terms, a CS researcher may store his/her documents in two types of storage means. If documents are physical, then they may be stored in shelf, drawers, desk tables or folders. Digital document, in turn, can be locally saved in a storage device —either in external devices, or in internal devices included in computers, laptops, smartphones or tablets—, or in a remote means, like basic cloud storage services —for example Dropbox, OneDrive, or Google Drive— that offer a virtual disk to store data, or more functional services, like those provided by some reference managers that offer classifying, tagging, or annotating the documents stored in the cloud —for example Mendeley or Refworks.

At last, each CS researcher can have some preferences when naming both the document files and the folders in which are stored these files. In general, CS researchers use the name of a document file to annotate the relevant information that can help them to easily recognize the document later. However, which aspect is considered relevant depends on each CS researcher: the document’s title —either the full title or part of it—, some keywords extracted from the document, some personal keywords, the name of the main author —or of the author considered more relevant by the CS researcher—, the publication year, etc. In the case of the folders, CS researchers follow the same approach and, on occasion, they reflect the purpose of the documents contained in it, like “to be read” or “for students of human-computer interaction course”.

5.1.4 History of CS researcher’s usage of the information-seeking system

In order to capture or infer all the preferences explained above, the system has to gather a great amount of information not only about the CS researcher, but also about the use he/she makes of the system in order to achieve his/her information-seeking purposes. In fact, in the context of adaptive systems like web-based systems, Kobsa, Koenemann, and Pohl (2001) has stated that the data used in the information-seeking activities have to be taken into account to model the process, considering not only the user data —that is those data related to the user’s characteristics—, but also environment data —that is those data that are external to the user, like those of the context explained in Section 3.6 of Chapter 3—, and usage data —that is those data about user interaction with the system. The history of usage presented in this section fits into the latter group.
In general, in any kind of software it is advisable to avoid asking the user, as long as possible, to provide information to the system as this may result in a reduction of his/her user experience. In fact, a system should be designed so that the user only has to perform those tasks that are relevant for him/her, instead of having to carry out system tasks that are needed for personalization and configuration purposes, even if this improves the performance of the system. In an interactive software system, this is even more important, as carrying out tasks that are not centered on the user can result in breaking the action flow of the user, and then in significantly affecting the usability of the software system and the user experience felt by the user.

It is why in this solution a machine learning approach is proposed, in order to allow tracking all the actions performed by the CS researcher, and then, transparently to the him/her, to process all this information and infer which are his/her main preferences. This approach implies accepting a certain margin of error due to the inference process, especially at the beginning. This margin, however, should decrease in time as the CS researcher makes more use of the system, as more information is gathered and then the inference is more robust and can interpolate more credible, more trustworthy and more in line with reality results.

Nevertheless, there are some aspects that are outside the system that cannot be tracked, or that lack enough information to infer a statement, and then it is inevitable to have to ask the user. This, for example, occurs when identifying which information-seeking purpose has the CS researcher, as the number and type of tasks that are performed by the CS researcher, and the order in which they are performed cannot be known by the system, at least, until the information-seeking purpose has been achieved. Nonetheless, if the system is able to model how the CS researcher usually carries out the information-seeking activity depending on his/her purpose, then it can anticipate to his/her needs once he/she has indicated the information-seeking purpose to the system. Analogously, there are some aspects related to the context that can neither be inferred by the system, and then have to be explicitly specified by the CS researcher, like the environment in which he/she is, the amount of time he/she has available to carry out the activity, the type of work style, or the use of physical work areas.

However, most of the preferences can be obtained automatically without needing to ask the user to provide any specific information. To do this, it is required to track the CS researcher’s actions and store as much information related to them as possible. Figure 5.11 reflects four types of actions that should be logged in the CS researcher’s history of usage of the information visualization system. Following sub-sections define these actions.
Figure 5.11 Actions recorded in the information-seeking history.
5.1 Main components of the researcher’s model

5.1.4.0.1 System-related actions
In this category are included all the general actions performed by the CS researcher that are directly related with the information system itself, like saving the current state of an information-seeking activity, or loading a previously saved state. In both cases, the amount of times the state of an activity is saved and then loaded indicates if the CS researcher does it at once, or if he/she needs to resume the activity several times, and then which is the amount time he/she usually has available to carry out an information-seeking activity.

Additionally, recording the context of the activity both when the CS researcher saves or loads its state can help the system recognizing when the CS researcher usually starts—or resumes—the activity and how much time he/she usually spends on it before saving it, but also where is the activity performed, both at the workspace level—for example, at home or at the office—, and at the device level—for example in a desktop computer or in a tablet.

Similarly, storing which documents are still in the visualization and how they are clustered and filtered when the state is saved can give some information about which are the checkpoints used by the CS researcher to progressively refine a collection to achieve an information-seeking purpose.

In this category can also be included all the actions aimed to set up the system, like selecting the dimensionality of the visualization workspace, indicating if the legend has to be displayed—and if so, where and how—, the setup of the workspace—for example the amount and type of devices used—, or which means of access he/she wants to use to obtain a collection of documents—for example local or remote, and in the latter case which one(s).

5.1.4.0.2 Document-related actions
In an information-seeking activity, everything revolves about how the CS researcher visualizes and manipulates the documents, and how he/she uses them. Then, it is essential to capture all the actions performed by the CS researcher over the documents and their representations to properly model many of the aspects related to the CS researcher. These actions, in turn, can be divided into two main categories, depending on the target of the action:

a) Actions performed on the document itself

A document can be saved in a storage means. In this case, the system has to record where the document—identified with a unique identifier—has been stored, that is in which type of storage means—for example, in which device or in which remote means—, with which name—for example, title of the document, and/or publication
year, and/or main author, etc.—, and in which path —especially the name of the folder in which it is stored.

**Renaming** a document already stored in a storage means is also an action that has to be stored. In this case, it is important to know the initial name of the file, and which is the new name assigned by the CS researcher. Collecting this information is essential in order to model the storage & management preferences of the CS researcher, namely his/her preferred naming strategy, storage means and storage classification strategy.

A CS researcher can also decide to **delete** a document from the storage means. In this case, knowing which document has been deleted is very relevant so that the system can indicate the CS researcher that the document has been deleted if it appears as a result in future searches.

The system has to track if the CS researcher **prints** a document, as this can indicate that he/she prefers to read it in paper format. Additionally, it may be useful to keep track of what documents have been printed, so that the system can indicate it to the CS researcher in case, for example, he/she has taken manual notes on the printed version.

Similarly, if a CS researcher **opens** a document, the system has to log it. If possible, the system should be able to track which parts reads the CS researcher and how much time he/she spends on each of them, so that it can identify which are the most important sections for him/her.

After entirely or partly reading a document, the CS researcher may want to indicate how **relevant** a document is for him/her, and may find useful to **add some notes** to it whether to summarize which are the most relevant information contained in the document, or to indicate some ideas related to the document he/she may have, for example for future works. Registering this information can help the system identifying, first of all, if the CS researcher likes adding notes and indicating the relevance of a document, and which are the documents implied in these actions. In both cases, the system has to record the value of the action —that is the note itself in the first case, or the relevance degree in the second case. In the case of the relevance, the system has also to record if the CS researcher changes the relevance of a document, as this may reflect, for example, that the document has become obsolete if, after being considered relevant for some time, the CS researcher considers to decrease its relevance.

Finally, as a CS researcher may be interested in different topics, it seems reasonable that he/she has different collections of documents, differentiated according to the
main components of the researcher’s model

5.1 Main components of the researcher’s model

topic(s) they address. Then, recording to which collection(s) is associated a document gives a lot of information about the topics in which the CS researcher is interested, as he/she is the one providing a first high-level clustering.

b) Actions that affect the content of the document

While reading a document, a CS researcher may consider part of the document very relevant, and may want to highlight it so that he/she can easily find it again in future readings. This, for example, can help to establish which are the terms and content elements that are considered relevant by the CS researcher in order to define his/her topics of interest.

On the contrary, it is also interesting to register which parts of the content are crossed out by the CS researcher, as this indicates in which contents and terms he/she is not interested.

Moreover, the CS researcher may want to add some comments to a given sentence or figure in the document. As before, capturing these comments, and to which content element they are associated, can be used by the system to identify relevant terms and to help the user processing more quickly the document in future readings.

5.1.4.0.3 Visualization-related actions

a) Actions performed on the document’s representation

As explained in Chapter 4 one of the main types of visual representations is the one illustrating the documents in the visualization. Here are presented the actions whose target is the document’s representation that have to be logged in the information-seeking history.

The most basic action that can be performed on a document representation is to select it. In fact, this action is simple but essential, as all the actions that are somehow related to one or more documents require first to define to which documents they are applied, and this is done through selection. This selection may be done directly by the CS researcher, who may directly select the representation —for example by clicking over it—, or indirectly through the use of another element, like a list of documents.

Additionally, the selection may be individual but also several documents may be selected. In this second case, the selection can also be made manually —for example using a selection area— or by using another element, like a list, which in this case has also to reflect groups. Generally, these groups would stem from a previous
clustering task. If so, the system has to record which is the criterion—or criteria—that define the selected cluster of documents. In any case, the system has to record which documents have been selected.

Having all this information allows the system inferring how relevant a document is based on the use the CS researcher has made of it. For example, if a document is regularly selected to consult some its information, it may be inferred that it has a high relevance for the user. More details about the relevance calculation based on the CS researcher’s history of use are provided in Chapter 6.

One of the actions that may be performed after selecting a document representation is to **clear it from the visualization**. In this case, again, one or more document representations can be manually cleared at a time, depending on how many documents have been previously selected.

However, document representations can be also cleared from the visualization without being previously selected if, for example, a complete previously created cluster wants to be hidden. In this case, again, the system has to log the criterion—or criteria—that define the affected cluster, and the identifier of the documents cleared from the visualization.

Using this information, the system can infer if the CS researcher prefers filtering the collections manually or through some filtering mechanisms.

The last type of action that can affect the visibility of the document representations is the execution of a filtering task. In this case, those documents that do not match the filtering criteria—if an inclusive strategy is followed—are cleared from visualization. In these cases, the system has to record which filtering mechanism has been used, namely the query and/or the filtering constraints. In the first case, the system has to log which are the terms used in the query, as they probably are key terms for the CS researcher, and then they are useful to model his/her topics of interest. In the second case, the system has to know which are the most common information elements used to filter a document collection, so that it can suggest its use to the user in future filtering tasks.

In this kind of actions, it is also important to keep track of how many documents are affected by the filtering, and which are they. This information can allow the system inferring if the CS researcher prefers to obtain precise or exhaustive results, and what’s the CS researcher’s preferred magnitude to stop filtering the collection.

However, before selecting or clearing from the visualization all the documents belonging to a cluster, the **visual clustering** has to be performed. If so, the system has to record all the data leading to the creation of the clusters. This implies
registering, first of all, which is the clustering criterion used, namely the information element used as clustering element, and its value in case a binary clustering wants to be obtained—for instance, dividing a document collection in two groups: those documents published before a given date, and those published on this date or later. This information allows the system determining which are the information elements preferred by the user to explore a document collection. Apart from this, clusters have to be visually represented, and this implies modifying their visualization dimensions, as explained in Chapter 4. As each CS researcher may prefer to use one or another visualization dimension, it is important for the system to identify which are the most common visualization dimensions used by the CS researcher. Additionally, as each visualization dimension is used to illustrate a given clustering criterion based on an information element, the system has also to register which is the association made by the CS researcher, as it can reflect one of his/her visualization-related preferences. For example, a CS researcher can like to visualize the impact metric of the documents through the size, while another may prefer the system to use the color saturation. Then, if the system knows this preference, it can directly suggest it to the CS researcher.

In the same way, in Chapter 4 is explained that there are many visual representations that can be used in an information visualizations system, and each of them can be illustrated according to a different modality. Also in this case, the system has to register which are the modalities that are more commonly selected by the CS researcher to illustrate, for example, the documents or the legend, in order to be able to automatically adapt the visualization to these preferences.

Finally, regardless the dimensionality of the environment, the system may allow the user to translate the document representations within the visualization workspace—both in a list, in a plane or in the space. If so, it has to keep track of the movements made, registering which documents are moved, and from where to where. The system has to know these modifications as, if for example the position in a list reflects the relevance of the documents, or the position in the space reflects the thematic similitude between documents, and the CS researcher changes the position of a document representation, it then means that the system has to change the relevance or similitude or any other aspect of the document to ensure it internally represents what the CS researcher has visually indicated.

b) Actions performed on the visualization’s point of view
In case a two-dimensional or three-dimensional environment is used to visualize the documents collections, it is essential to provide the CS researcher with mechanisms
Model of Computer Science researchers

to fully explore the environment. This implies allowing to translate and rotate the virtual camera, as CS researcher sees the virtual world from its point of view. In both cases, having this information can allow the system to determine, for example, which is the magnitude of the visualization focus when the CS researcher is exploring a set of documents.

**Translating** the view in the horizontal and vertical axes allow the CS researcher navigating within the virtual world in order to, for example, center the visualization focus on a given area where representations of relevant documents are located, or just to be able to visualize a document representation located outside the current visualization focus. If the third dimension is also used, then the CS researcher has to be able to **zoom in and out** to perceive the digital environment at different levels of depth. In this case, the system has to record all the translations, that is the axis along which the camera moves, the direction of the movement in this axis, and the distance traveled. Alternatively, this values can be calculated if the initial and final positions of the camera are recorded. In the case of the zoom, it is more interesting to know the zoom level applied.

Similarly, if the CS researcher can **rotate** the view to observe the visualization workspace from a different orientation, the system has to keep track of these rotations—around which axis and how many degrees, or the initial and final orientation.

### 5.2 CS Researcher’s topics of interest

One of the aspects that is essential when modeling a CS researcher is the identification of the topics in which he/she is interested. This implies, on the one hand, specifying how the topic itself is defined, and on the other one, how the system can determine if a topic is of interest or not for a given CS researcher.

The approach followed in this research is based on the user modeling proposed by Ahn and Brusilovsky (2013), based in a two-steps process. In their proposal, first of all, the user manually selects text fragments from the documents returned by a search engine in order to create a task model that represents his/her current interests; then, these fragment are processed in order to extract relevant terms, that are finally used to update the user’s model. In this research a similar approach has been adopted, but applying it only to identify which are the terms that are more relevant for a CS researcher. The purpose is to model the interests of the CS researcher through a vector in the VSM that contains the frequency of these terms, as done for the topic modeling—see Chapter 3.
5.2 CS Researcher’s topics of interest

In order to improve the calculation of the vector of the CS researcher’s topics of interest, the proposed solution goes a little bit beyond the solution proposed by Ahn and Brusilovsky as not only the set of terms explicitly indicated by the CS researcher are used, but also those that are included in the documents that are considered relevant by the CS researcher, in the documents written by authors that are considered relevant by the CS researcher, and in the documents published in venues considered relevant by the CS researcher. How to determine which of these aspects are considered relevant or not by a CS researcher, and to what extent they are relevant for him/her, is explained in Chapter 6.

Apart from this, in order to properly identify which are the terms that are actually relevant for the CS researcher and then make up the user model, it has to be considered to which degree the source from where they come is representative of a topic. To do so, we propose weighting the influence of the terms according to the source from where they come. The following scheme is adopted:

- terms explicitly indicated by the CS researcher must have a predominant weight, as they reflect his/her conscious interests and/or relevant aspects;

- terms appearing in a document considered relevant by the CS researcher also have to have an important influence, as documents usually address very few topics—usually one—and then it seems reasonable to consider that the terms included in it can also be considered as relevant to model the CS researcher’s interest;

- terms contained in documents written by an author who is relevant for the CS researcher has to have a limited impact when defining the CS researcher’s topics of interest, as an author can be interested in many topics, that may be somehow related or not, and then it cannot be assumed that all documents written by an author address only the topic that the CS researcher considers relevant. By extension, then, the terms contained in these documents can neither be considered as totally relevant for a given topic;

- terms contained in documents published in a publication venue considered relevant by the CS researcher may be considered but with great caution, as journals and conference have wide aims and scopes, and then they can address many topics. This makes that most of the terms that may be considered relevant for other topics are probably going to distort the modeling of the CS researcher’s topics of interests. However, even if this weighting greatly limits its influence, considering these terms can provide interesting nuances when modeling the user’s topics of interest.
Then, a decreasing weighting scheme, illustrated in Figure 5.12, is proposed so that the weight of a term is indirectly proportional to the degree of correlation between its provenance and the topics of interest of the CS researcher:

- user’s explicit terms have a weight of 1;
- terms from relevant documents have a weight of 0.5;
- terms from documents written by relevant authors have a weight of 0.25;
- terms from documents published in relevant publication venues have a weight of 0.1.

Figure 5.12 Weighted vector representing the relevant terms for the CS researcher

As the relevance of a term is dependent on its frequency of occurrence, this means that, for example, a term appearing in a document that has been published in a relevant publication venue would have to appear ten more times than a term explicitly indicated by the CS researcher to have the same influence.

Finally, to determine to what extent a document addresses the CS researcher’s topics of interest, it is enough to calculate the similarity metric between the vectors that represent them in the VSM, as explained in Section 3.4.3 of Chapter 3.
Chapter 6

Adaptive relevance estimation: A Fuzzy Logic approach

As stated in Chapter 5, one of the aspects that defines a CS researcher is the amount of relevant knowledge the researcher has. Part of this knowledge is related to the information-seeking activities themselves, while the other part is related to a topic like who are the reference authors, which are the reference research groups that do research on it, which are the reference documents and/or publication venues dealing with it, or what its specific terminology is. The relevance of this knowledge cannot be defined in general terms, as it is clearly linked to a specific topic, but it can neither be associated only to the topic. In fact, relevance is also dependent on the researcher and on the information-seeking purpose, basically for two reasons. First of all, defining an author, for example, as very relevant in a topic without considering any other aspect would imply that the author is a reference for anybody researching in this topic, no matter what are his/her specific interests. However, this is not always true, as it may occur that the specific research interests of a researcher only cover a small subpart of the topic, and then the relevant information related to the rest of the topic is actually not relevant for the researcher. In the same way, it is entirely feasible that a researcher is interested in different research fields that can be related to each other, or not. If they are related, it is possible that, in varying degrees, the relevant knowledge is the same. However, if the fields are not related in any way, it would be a great mistake to mix their respective relevant knowledge. As an example, a researcher may be interested in virtual reality and in information visualization and, even if some authors or documents can address both fields at the same time, there probably are a lot of other knowledge that is relevant in only one or the other field, but not in both of them. This means, then, that the solution has to be able to reflect that the same researcher can have different profiles in terms of topics of interest, which in this case is done by relating
the relevant knowledge both to the researcher and to the topic, as reflected in the CS researcher’s model —see Chapter 5.

Acquiring this knowledge is essential in order to provide adaptive capacity to the system as it 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 exploration. Indeed, in any research topic, there are many researchers that can be considered experts, either because they are senior researchers and then are proficient undertaking an information-seeking activity, or because they have a lot of knowledge about certain topics. It is important to distinguish interest, expertise and relevance. A researcher in general can be interested in a very large range of topics, but he/she probably is an expert in only some of them —like a Ph.D. supervisor, a university professor— due to the large amount of knowledge acquired and consolidated over time. Evenly, expertise does not mean relevance, as a researcher can have a lot of knowledge about a topic, but this does not imply that he/she is a relevant author in this topic because, for example, he/she may have not produced any relevant contribution to the scientific community and then he/she may not be considered a reference author in the topic.

In this work it has been determined that there are five entities that form the knowledge a CS researcher can have about a given research topic —documents, authors, publication venues, research groups and specific terminology. In order to calculate the relevance of these entities, a set of numerical counters registering, for example, the amount of times a CS researcher has opened a given document could be used. Then, the relevance could be figured out by applying a set of rules using thresholds to define the extent of the relevance. As an example, a rule could be: "If the user opens a document more than 10 times, then the document is very relevant for him/her". However, this rule is very strict and then does not faithfully illustrates a real situation. In the previous example, opening 9 times a document should also indicate that the document is very relevant, even if a little bit less than the one opened 10 times. On the other, it would be unrepresentative to map very abstract and vague expressions like "somewhat relevant" or "very relevant" to simple numbers.

To manage this vagueness and variability in the decision-making process embodied in the rules, and to allow a more flexible and realistic approach to calculate an abstract aspect like the relevance of an entity in the research field, fuzzy logic is used, as it perfectly handles these aspects. Fuzzy logic perfectly adapts to the thinking of human beings and allows to reliable reflect their complex mental models, as it makes use of natural language expressions to define the rules, which makes them very close to the user. This is essential to design adaptive interactive systems, as the ones we propose in
6.1 Fuzzy logic

this thesis, as the system is able to "speak" the same language as the user and then prevents the users to have specific knowledge about more complex mathematic models (Zadeh, 1973). Thanks to this, the user finds easier and more intuitive to interact with the system, which greatly improves his/her user experience.

An existing Java library called jFuzzyLogic (Cingolani & Alcalá-Fdez, 2012; Cingolani & Alcala-Fdez, 2013) that allows implementing and using a fuzzy inference system in a Java environment has been used in this work. The implementation adheres to the standard for the Fuzzy Control Language (FCL) proposed by the International Electrotechnical Commission (IEC) (IEC, 2000).

In order to calculate the relevance of the mentioned entities, three aspects have been considered in his work:

1. **Objective relevance.** The relevance of the entities been assessed —that is authors, documents, publication venues and research groups— has to have into consideration their objective relevance, that is the one base don objective indicators and that is independent from the researcher who uses it or the purpose for what it is used. This component of the relevance is explained in more detail in Section 6.2.

2. **Researcher-dependent relevance.** In this case, it is considered the relevance a specific researcher assigns to a given entity based on his/her interests, preferences and purposes. This second component is developped in Section 6.3.

3. **Inter-concepts influence.** The third component of the relevance of an entity in the research field is the one deriving from the relevance of other entities that are related to the assessed one. More details are provided in Section 6.4.

Finally, in Section 6.5, an integration of these three components is proposed in order to calculate holistically the final relevance of an entity.

In this case, the fuzzy system proposed to estimate the relevance of an entity in the research field is not dependent on the user profile as the aspects considered in the system are domain-dependent —research field—, but not user-dependent. This means that the proposed approach can be applied to any researcher, not only to CS researchers.

### 6.1 Fuzzy logic

The basic principle of fuzzy logic consists in translating numerical raw data into fuzzy data which, after being operated between them, returned a fuzzy result that is in
Adaptive relevance estimation: A Fuzzy Logic approach

turn translated into a numerical value (Tanaka, 1997). The strength of this approach lies in the fact that fuzzification is performed according to a set of functions—called membership functions—that taking the numerical value of a variable as input return the membership level of this variable to a semantical label—called linguistic term. This membership level indicates with what certainty it can be affirmed that a variable has a given linguistic term.

Let’s take as example the variable age. In classical logic, it would be difficult to say if a person is young or old, as it is a vague concept. In fuzzy logic, on the contrary, two linguistic terms can be defined to represent how old is a person using natural language: 'young' and 'old'—see Figure 6.1. The linguistic term 'young' could be defined by the linear membership function $y = 1 - x/60$, and 'old' by $y = x/60 - 2/3$, being $y \in [0, 1]$. In this case, then:

- a 20 years old person would be 'young' with a membership value of 0.667, and 'old' with 0 as membership value;
- a 85 years old person would be 'young' with a membership value of 0, while he/she would be 'old' with a membership value of 0.75;
- a 55 years old person would be 'young' with a membership value of 0.083 and 'old' with a membership value of 0.25.

![Figure 6.1 Example of fuzzy functions for the linguistic terms 'young' and 'old' of the variable 'age'](image)

In the two first cases, where the input value affects only one linguistic term, the analysis is simple: a person is mainly young or old. However, the analysis of the age of the third person is more interesting and a little bit more complex as the variable is
fuzzified into two linguistic terms. In this case, the person can be considered a little bit old, while still slightly young. This type of analysis is possible due to the use of fuzzy logic.

Fuzzy logic also allows defining a set of rules where the values of the linguistic terms associated to the input variables are used as antecedents, and the consequents contain the values of the linguistic terms of the output variable. Within each rule, several linguistic terms with different values can be used as antecedent, and then have to be combined to determine if the consequent of the rule is triggered or not. This combination—also called Aggregation—is mainly based on the basic logical operators \( \text{AND} \) —Intersection—, \( \text{OR} \) —Union—, and \( \text{NOT} \) —Complement.

Moreover, there may be several rules with the same output variable as consequent, and then a single output variable can have assigned different linguistic term values, each of them with a different membership level. Then, in order to provide a unique value of the output variable, the different consequents have to be combined. This is done by operating the areas—called fuzzy sets—delimited by the membership functions of the linguistic terms intervening in the consequents of the rules—the operation is called Accumulation. Some of the most common accumulation methods are Maximum, Bound Sum, and Normalized Sum—see (IEC, 2000) for more details. At the end, then, the output fuzzy variable is represented by the area resulting from the integration of the fuzzy sets defined in the consequents of the activated rules.

Let’s consider the input variables "height" and "experience" and the output variable "quality_player". Each of these variables is defined by three linguistic terms: "height"—tall, normal, small—, "experience"—few, some, many—and "quality_player"—bad, normal, good. Now, let’s pose a couple of rules:

- **IF** experience IS some **THEN** quality_player IS normal
- **IF** experience IS some **AND** height IS tall **THEN** quality_player IS good

In this case, for example, if the player has some experience and is tall, then both rules get activated and then their corresponding areas have to be integrated in order to determine his/her quality.

Additionally, fuzzy logic allows differentiating the weight each rule has to have when calculating the output fuzzy variable value. This is very useful as it allows defining how much influence has to have a given rule on the result. In jFuzzyLogic, this weighting is indicated by adding the clause "\( \text{WITH} \) weight" at the end of the rule.

The last step consists in defuzzifying the output variable so that the area of aggregated fuzzy set—resulting from the aggregation of all the areas delimited by the membership functions derived from the consequents of the activated rules—is returned.
as a numerical value. To do so, there are some defuzzification methods that take as input this area and return a numerical value, like Center of Gravity, Center of Area, or Mean Max—see (IEC, 2000) for more details.

In this solution, the accumulation method Maximum is the one that has been used, and Center of Gravity as defuzzification method, as they both are the most commonly used.

6.2 Objective relevance based on indicators

6.2.1 Document’s relevance based on the amount of times it has been cited

One of the main criteria used to determine how relevant is a document is the amount of times it has been cited by other documents. In general, a researcher can consider a document relevant if it has been cited many times, as this implies that many other researchers, after reading it, have considered it relevant enough to be used as a reference in their works. There are two approaches that can be followed to calculate this component of the document’s relevance:

- the system can make use of the total number of citations, like the one provided, for example, by Google Scholar,
- the system can calculate the average number of citations a document has had during a certain period, like the one provided, for example, by Thomson Reuters in their annual Journal Citation Report.

In order to provide a wider view of the document’s usage, the first option has been used. In terms of fuzzy logic, this aspect can be seen as an input variable used to calculate the target output variable—objective document’s relevance—, and then it has to be fuzzified.

However, it is not obvious determining, even with the flexibility provided by fuzzy logic, when a document has been little, quite or very cited as it depends on other aspects, like the specific research field it addresses. As an example, a document published in the medicine research field requires a lot more citations than a document in the anthropology research field to be considered a much cited document. This also happens with the document type: a relevant book usually has a lot more citations than a relevant document, as the first one is supposed to more broadly cover the topic and to present very believable and contrasted information, and then it is easier for it to be cited as basic related work in many of the research documents addressing the same topic. As
this work focuses on the research in computer science, the magnitude of citations that relevant documents have —and more specifically relevant articles— in this area is the option that has been used as reference.

Additionally, the information-seeking purpose of the researcher can also determine when a document can be considered quite or very cited. As an example, if the purpose is to find a reference to support a given statement, it is maybe not so important for the researcher to find a very cited document, whereas if he/she is elaborating the state-of-the-art of the topic, it is essential to find and process all the documents that have been widely cited in the literature, as avoiding them could result in an incomplete analysis of the related works.

Another interesting point that should be considered when determining the relevance of a document based on the number of times it has been cited, is the own relevance of the citing documents. In fact, a document cited by two extremely relevant documents should be much more relevant than another one cited by five irrelevant documents. In our opinion, if a document is referred by a relevant document, it is likely for the first one to be also somewhat relevant, as the confidence that the relevant document generates to the researcher makes him/her trust on the good judgment of its authors when selecting their reference. Additionally, the finding of related relevant documents gets easier when the task is distributed among many researchers.

In this sense, it could also be discussed if the documents citing a relevant document should also be considered somewhat relevant. In this work, it has been considered that citing a reference document does not ensure any kind of relevance, as it does not imply that the contribution of the citing document is interesting, relevant and of good quality, but only that the authors have carried out a good literature review. It could be also discussed if, for the sake of allowing the researcher discovering documents that are unknown for him/her and that may be of interest, the system should assign a little relevance to them so that they are more noticeable by the researcher. In this work it has been considered that introducing this relevance would introduce more noise than benefits in the calculation of the document relevance, and then it has not been taken into account.

So, if the relevance of the citing documents is considered instead of the simple amount of citations, there are again two options that may be adopted:

- considering the objective relevance of the citing documents based on the amount of times they have been cited —using, for example, the value provided by Google Scholar—, or
- using the final relevance, that considers the three components of the relevance, so that not only the objective relevance is taken into consideration, but also
the subjective relevance that has the document for a given researcher, and the relevance propagated by the other concepts that are related with the document—like the its authors or its publication venue.

As the final aim is to design an adaptive system, the second option is more suitable as it takes into account the preferences of the researcher and the context of the document. As the final relevance of a document has still to be defined, the complete definition of this objective relevance is provided in Section 6.5.1.

6.2.2 Author’s relevance based on the amount of documents he/she has written

The relevance of an author can be partly calculated with the amount of documents he/she has written. In fact, it is more likely for a prolific author to be relevant in a topic, as this means that he/she has conducted many investigations and has produced a lot of results, than for an author generating sporadic and punctual results.

However, as it happens with the relevance of a document based on the amount of times it has been cited—already presented in Section 6.2.1—, quantity does not always mean quality or relevance, as an author generating less documents may be very relevant if these documents are a reference in the addressed topic. So, in this case the amount of documents an author has written is not the only aspect that is taken into account, but also the final relevance of these documents.

Then, the definition of this input fuzzy variable is also completed later, in Section 6.5.2, after explaining how the final relevance of an author is calculated.

6.2.3 Author’s relevance based on his/her impact metrics

As reflected in the proposed solution, the relevance of an author can be determined according to one or more impact metrics. As it happens with publication venues, basing the relevance of an author, at least partially, on one or more impact metrics has some limitations due to the inner restrictions of the methods used to calculate them, but on the contrary it has the advantage of making use of an element that has been externally calculated according to a well-known formula, and that is recognized and accepted in the research field, which brings credibility to the result. In this solution three impact metrics that are widely known and used in the research field are used:

- **h-index**: the metric, proposed by Hirsch (2005), indicates that the author has published $h$ documents that have been individually cited at least $h$ times.
6.2 Objective relevance based on indicators

- **g-index**: proposed by Egghe (2006) as an improvement of the h-index, it reflects that the top $g$ most cited articles of an author have received together at least $g^2$ citations. This metric, then, gives more value to highly-cited papers.

- **RG score**: this metric is provided by the research-oriented social network Research Gate\(^1\) to illustrate the scientific reputation of an author. This metric is highly controversial as it may not reflect the real relevance of an author, mainly because it does not systematically take into account all the work performed by the author, but only the one he/she has introduced in the network, and because it depends on his/her social contributions to the community, like participating in forum debates or answering questions from other researchers. Despite this, this metric has been included in the calculation of the relevance as it has this social dimension that the other metrics do not have, and it can provide a more holistic approach. However, in order to avoid miscalculating the relevance of an author based on his/her impact metrics due to unrealistics values, its influence in the calculation has been limited.

Next are presented the linguistics terms and membership functions used to define the fuzzy variables used to represent the author’s relevance based on the value of each of these impact metrics:

- **h-index**:
  
  - **low relevance** is calculated with the function: $(0, 1) (5, 1) (15, 0)$
  
  - **medium relevance** is calculated with the function: $(10, 0) (25, 1) (40, 0)$
  
  - **high relevance** is calculated with the function: $(30, 0) (50, 1) (70, 0)$
  
  - **very_high relevance** is calculated with the function: $(60, 0) (80, 1) (120, 1)$

- **g-index**:
  
  - **low relevance** is calculated with the function: $(0, 1) (15, 1) (30, 0)$
  
  - **medium relevance** is calculated with the function: $(20, 0) (55, 1) (90, 0)$
  
  - **high relevance** is calculated with the function: $(80, 0) (120, 1) (160, 0)$
  
  - **very_high relevance** is calculated with the function: $(150, 0) (205, 1) (250, 1)$

- **RG score**:
  
  - **low relevance** is calculated with the function: $(0, 1) (5, 1) (10, 0)$

\(^1\)https://www.researchgate.net/RGScore/FAQ
Adaptive relevance estimation: A Fuzzy Logic approach

- medium relevance is calculated with the function: (5, 0) (15, 1) (25, 0)
- high relevance is calculated with the function: (20, 0) (35, 1) (50, 0)
- very_high relevance is calculated with the function: (40, 0) (60, 1) (80, 1)

Figures 6.2a, 6.2b and 6.2c illustrate the definition of these variables in fuzzy logic.

(a) Variable "h-index of the author"
(b) Variable "g-index of the author"
(c) Variable "RG Score of the author"

Figure 6.2 Membership functions for the fuzzy input variables related to the author’s impact metrics

In order to integrate all these values into a single one that reflects the overall relevance of an author based on his/her impact metrics, it is needed to define a new output fuzzy variable, 'relevance of the author based on his/her impact metrics', and a set of rules allowing to calculate its values based on the membership levels of the input fuzzy variables defined just above.

Figure 6.3 illustrates this output variable, which is defined by the following linguistic terms:

- low relevance is calculated with the function: (0, 1) (0.2, 1) (0.3, 0)
- medium relevance is calculated with the function: (0.2, 0) (0.4, 1) (0.5, 1) (0.6, 0)
- high relevance is calculated with the function: (0.5, 0) (0.7, 1) (0.8, 1) (0.9, 0)
- very_high relevance is calculated with the function: (0.8, 0) (1, 1)
At this point, it is required to combine the results of all these input variables in order to obtain the value of the variable that actually needs to be calculated. In order to do so, fuzzy logic allows defining a set of IF...THEN rules in which both the antecedent and the consequent are defined by indicating which linguistic terms have to have the variables. Next are presented the rules used to calculate the mentioned output fuzzy variable:

- **Rule 1:**
  IF (h_index IS very_high) AND
  (g_index IS very_high OR g_index IS high)
  THEN author_relevance_by_impact IS very_high

- **Rule 2:**
  IF (h_index IS high) AND
  (g_index IS very_high)
  THEN author_relevance_by_impact IS very_high

- **Rule 3:**
  IF (h_index IS high) AND
  (g_index IS high) AND
  (rg_score IS very_high)
  THEN author_relevance_by_impact IS very_high WITH 0.7

- **Rule 4:**
  IF (h_index IS high) AND
  (g_index IS high) AND
  (rg_score IS NOT very_high)
THEN author_relevance_by_impact IS high

- Rule 5:
  IF (h_index IS high) AND
  (g_index IS medium) AND
  (rg_score IS very_high OR rg_score IS high)
  THEN author_relevance_by_impact IS very_high WITH 0.7

- Rule 6:
  IF (h_index IS medium) AND
  (g_index IS high) AND
  (rg_score IS very_high OR rg_score IS high)
  THEN author_relevance_by_impact IS very_high WITH 0.7

- Rule 7:
  IF (h_index IS medium) AND
  (g_index IS medium)
  THEN author_relevance_by_impact IS medium

- Rule 8:
  IF (h_index IS low) AND
  (g_index IS medium) AND
  (rg_score IS very_high OR rg_score IS high)
  THEN author_relevance_by_impact IS medium WITH 0.7

- Rule 9:
  IF (h_index IS medium) AND
  (g_index IS low) AND
  (rg_score IS very_high OR rg_score IS high)
  THEN author_relevance_by_impact IS medium WITH 0.7

- Rule 10:
  IF (h_index IS low) AND
  (g_index IS low)
  THEN author_relevance_by_impact IS low

- Rule 11:
  IF (h_index IS low OR g_index IS low) AND
(rg_score IS medium OR rg_score IS low)

THEN author_relevance_by_impact IS low WITH 0.7

As indicated before, because of the doubts concerning the RG score, all the rules where this metric is used as antecedent to calculate the output variable—that is in rules 3, 5, 6, 8, 9 and 11—have been weighted so that, if the rule gets activated, its influence on the output variable calculation is reduced to 70% of its real influence. In the case of rule 4, the weighting is not applied as the RG score is only used to distinguish this rule from rule 3.

In order to illustrate how the fuzzy output variable “relevance of the author based on its impact metrics” is calculated, metrics of some real authors have been retrieved\(^2\) from Google Scholar\(^3\), Publish or Perish\(^4\) and Research Gate\(^5\):

- Ben Shneiderman has the following metrics: h-index = 103, g-index = 250, rg_score = 35.25. The fuzzy system returns as center of gravity of the area illustrated in Figure 6.4 the value \(0.933\), and the following membership levels for the different linguistic terms:
  - Low relevance: 0.0
  - Medium relevance: 0.0
  - High relevance: 0.0
  - Very high relevance: 0.665

- Peter Brusilovsky has the following metrics: h-index = 63, g-index = 143, rg_score = 31.57. The fuzzy system returns as center of gravity of the area illustrated in Figure 6.5 the value \(0.736\), and the following membership levels for the different linguistic terms:
  - Low relevance: 0.0
  - Medium relevance: 0.0
  - High relevance: 1.0
  - Very high relevance: 0.0

---

\(^2\)Data retrieved on 2016-07-07
\(^3\)https://scholar.google.es/
\(^4\)http://www.harzing.com/pop.htm
\(^5\)https://www.researchgate.net/
• Jae-Wook Ahn has the following metrics: h-index = 12, g-index = 25, rg_score = 11.55. The fuzzy system returns as center of gravity of the area illustrated in Figure 6.6 the value 0.2388, and the following membership levels for the different linguistic terms:

– Low relevance: 0.6124
– Medium relevance: 0.1938
– High relevance: 0.0
– Very high relevance: 0.0

Then, with almost 67% of certainty, the fuzzy system determines that Shneiderman is a very high relevant author based on his impact metrics, whereas Brusilovsky is considered an high-relevant author with 100% of certainty. In the case of Ahn, the system considers him as low relevant with around 61% of certainty, but also moderately relevant with around 19% of certainty.

Later, the defuzzified value returned by this output fuzzy variable is going to be used as input value to calculate the total relevance of authors —see Section 6.5.2. To do so, it is just needed to define the input fuzzy variable 'relevance of an author based on his/her impact metrics' just like this output variable —that is with the same membership functions assigned to the same linguistic terms. In this case, then, the value used as input value for this input variable would be the value returned by the fuzzy system after defuzzifying the output variable —that is 0.933, 0.736 and 0.2388 in the examples provided below.

6.2.4 Publication venue’s relevance based on its impact metrics

As mentioned before, the relevance of a publication venue is basically given by the value of one or more impact metrics, like Journal Citation Report (JCR)\(^6\), the SCImago Journal Rank (SJR)\(^7\), and the Latindex index\(^8\) in the case of journals, or the Conference Research & Education (CORE) ranking\(^9\) for conferences. Among these metrics, there are two types of values:

---

\(^6\)http://ipsience.thomsonreuters.com/product/journal-citation-reports
\(^7\)http://www.scimagojr.com/
\(^8\)http://www.latindex.org/latindex/inicio?lang=en
\(^9\)http://portal.core.edu.au/conf-ranks/
6.2 Objective relevance based on indicators

- Numerical values, like the ones given by the JCR or the SJR, that provide as impact metric of a journal the average number of times that has been cited a document published in it in the previous two or three years. In SJR, the citations are also weighted based on the relevance of the publication venue of the documents from where stems the citation.

- Ranking or clustering of the publication venues, like the quartiles proposed by the JCR, or the groups of the CORE.

In both cases, the impact metric indirectly makes reference to the relevance of the documents that have been published in the venue, limited to a specific period of time and taking into account only the number of times the published document has been cited. This means that the same publication venue has several values for the same variable, and then there are two ways to integrate them:

- only the newest value, corresponding to the last period of time analyzed, or
- as for research groups, it may not be a time restrictions, and then the average relevance can be calculated. Additionally, the result can be calculated by weighting the influence of the values according to their age, so that the newest metrics have more influence than the oldest ones.

In our opinion, the best option is the first one as the newest impact metrics’ values reflect the current relevance of the publication venue, which in general is the one in which the researcher is interested, for example, to select where to publish a new contribution.

6.2.4.1 Journal’s relevance based on its impact metrics

In the case of journals, numerical and classification impact metrics’ values provided by Thomson Reuters in its Journal Citation Report are combined to calculate the relevance of a publication, as they are two of the most common and used variables in the scientific research field. The numerical variable is defined in fuzzy logic as it follows:

- low impact factor is defined by function: \((0, 0)(0.5, 1)(0.75, 1)(1, 0)\)
- medium impact factor is defined by function: \((0.75, 0)(1.25, 1)(1.5, 1)(1.75, 0)\);
- high impact factor is defined by function: \((1.5, 0)(2, 1)(3, 1)(3.5, 0)\);
- very high impact factor is defined by function: \((3, 0)(4, 1)(\infty, 1)\)

The classification variable is in turn defined by the following linguistic terms:

- fourth quartile is represented by the value 4
Adaptive relevance estimation: A Fuzzy Logic approach

- *third quartile* is represented by the value 3
- *second quartile* is represented by the value 2
- *first quartile* is represented by the value 1

Figures 6.7a and 6.7b illustrate the definition of these variables in fuzzy logic.

![Figure 6.7 Membership functions for the input variables related to impact metrics of a journal](image)

(a) Variable "impact factor of the journal"
(b) Variable "quartile in which is clustered the journal"

Figure 6.7 Membership functions for the input variables related to impact metrics of a journal

These two input fuzzy variables, then, have to be combined under the output fuzzy variable "relevance of the publication venue based on its impact metrics", whose associated membership function is illustrated in Figure 6.8, and that is defined as follows:

- *low relevance* is calculated with the function: (0, 1) (0.2, 1) (0.3, 0)
- *medium relevance* is calculated with the function: (0.2, 0) (0.4, 1) (0.5, 1) (0.6, 0)
- *high relevance* is calculated with the function: (0.5, 0) (0.7, 1) (0.8, 1) (0.9, 0)
- *very_high relevance* is calculated with the function: (0.8, 0) (0.9, 1) (1, 1)

![Figure 6.8 Membership functions for the output variable "relevance of a journal based on its impact metrics"](image)

Figure 6.8 Membership functions for the output variable "relevance of a journal based on its impact metrics"

Finally, these are the rules defining the calculation of this output fuzzy variable when the venue is a journal:
• Rule 1:
  IF (quartile_journal IS first)
  THEN pubvenue_relevance_by_impact IS very_high

• Rule 2:
  IF (impact_factor_journal IS very_high) AND
  (quartile_journal IS second)
  THEN pubvenue_relevance_by_impact IS very_high

• Rule 3:
  IF (quartile_journal IS second)
  THEN pubvenue_relevance_by_impact IS high

• Rule 4:
  IF (impact_factor_journal IS very_high OR impact_factor_journal IS high)
  AND
  (quartile_journal IS NOT fourth)
  THEN pubvenue_relevance_by_impact IS high

• Rule 5:
  IF (quartile_journal IS third)
  THEN pubvenue_relevance_by_impact IS medium

• Rule 6:
  IF (impact_factor_journal IS very_high OR impact_factor_journal IS high)
  AND
  (quartile_journal IS fourth)
  THEN pubvenue_relevance_by_impact IS medium

• Rule 7:
  IF (quartile_journal IS fourth)
  THEN pubvenue_relevance_by_impact IS low

• Rule 8:
  IF (impact_factor_journal IS low)
  THEN pubvenue_relevance_by_impact IS low

In order to illustrate how the fuzzy output variable “relevance of a journal based on its impact metrics” is calculated, some examples are provided:
In the last year, journal 1 has the following metrics: JCR impact metric = 2.3, JCR quartile = Q1. The fuzzy system returns as center of gravity of the function illustrated in Figure 6.9 the value 0.922, and the following membership levels for the different linguistic terms:

- Low relevance: 0.0
- Medium relevance: 0.0
- High relevance: 0.0
- Very high relevance: 1.0

Let’s note that, if the same journal was classified in quartile Q2 instead of Q1, it would still be relevant, but to a lesser extent. As illustrated in Figure 6.10, its membership value would be 0.72, and the membership levels of its linguistic terms would be:

- Low relevance: 0.0
- Medium relevance: 1.0
- High relevance: 0.0
- Very high relevance: 0.0

Journal 2 has the following metrics: JCR impact metric = 1.2, JCR quartile = Q2. The fuzzy system returns as center of gravity of the function illustrated in Figure 6.11 the value 0.42, and the following membership levels for the different linguistic terms:

- Low relevance: 0.0
- Medium relevance: 1.0
- High relevance: 0.0
- Very high relevance: 0.0
6.2 Objective relevance based on indicators

- Journal 3 has an JCR impact metric 0.53 and is classified in Q4 (JCR quartile). The fuzzy system returns as center of gravity of the function illustrated in Figure 6.12 the value 0.1269, and the following membership levels for the different linguistic terms:
  - *Low relevance*: 1.0
  - *Medium relevance*: 0.0
  - *High relevance*: 0.0
  - *Very high relevance*: 0.0

As for the output fuzzy variable "author’s relevance based on its impact metrics" presented in Section 6.2.3, the output fuzzy variable “relevance of a journal based on its impact metrics” is going to be used as input fuzzy value to calculate the final relevance of publication venues —see Section 6.5.3. Then, the input fuzzy variable has to be define exactly like this output fuzzy variable, and the value returned by the fuzzy system has to be used as input value for the input fuzzy variable —in this case, these values are 0.922, 0.72, 0.42 and 0.1269.

6.2.4.2 Conference’s relevance based on its impact metrics

In the case of conference proceedings, as the classification proposed by the CORE ranking is the only metric used to calculate the conference relevance, and that it is a discrete metric that can only provide four values —AA (or A*), A, B, and C—, it is only needed to recode these values into numerical values, so that each of them exactly maps with the highest membership level of each of the linguistic terms of the fuzzy variable "relevance of a journal based on its impact metrics":

- ‘AA’ (‘A*’) is recoded with the input value 1
- ‘A’ is recoded with the input value 0.8
- ‘B’ is recoded with the input value 0.5
- ‘C’ is recoded with the input value 0.2

6.2.5 Relevance of a publication venue based on the amount of documents it has published

As in previous cases, it can be considered that a venue that has published many documents is more relevant than another less prolific one. However, as it happens with
the relevance of a document based on the amount of times it has been cited—see Section 6.2.1— and the relevance of an author based on the number of document he/she has written—see Section 6.2.2—, quantity does not always mean quality or relevance, as a very relevant publication venue can published very few documents per year, which allows it to accept only those documents that are of the highest quality, whereas a low-relevant venue may have to be more permissive and accept more documents even if their quality is lower. For that reason, as in the other cases, the final relevance of the documents that have been published in a given venue is taken into account instead of plainly using the number of published document.

Then, the definition of this input fuzzy variable is also completed later, in Section 6.5.3, after explaining how the final relevance of a document is calculated.

6.2.6 Relevance of a research group based on the amount of authors affiliated to it

As a research group is just a set of authors that work together and investigate related topics, the relevance of a specific group can be partly determined by the number of authors that are affiliated to it. In fact, it seems reasonable to state that a research group formed by many authors is probable more relevant than another one composed only by two or three members.

However, as it happens with the relevance of a document based on the amount of times it has been cited—see Section 6.2.1—, the relevance of an author based on the number of document he/she has written—see Section 6.2.2—, or the relevance of a publication venue based on the amount of documents it has published—see Section 6.2.5—, quantity does not always mean quality or relevance, as a group formed by three very relevant and prolific authors is almost certainly much more relevant than a group composed by ten authors who produce very few and low-quality research results. Then, again in this case, the amount of authors that are affiliated to a research group is not the only aspect that is taken into account, but also the final relevance of these authors.

Then, the definition of this input fuzzy variable is also completed later, in Section 6.5.4, after explaining how the final relevance of an author is calculated.
6.3 Researcher-dependent relevance

6.3.1 Document’s relevance based on the usage history of it by the researcher

As explained in Chapter 3, the document is the nuclear entity in information-seeking, as most of the other entities are related to it: authors write document, venues publish documents, and the specific terminology of the topics is extracted from documents. In fact, as it has been also explained in Chapter 4, in research-oriented information visualization systems, the main visual representation is the one illustrating the documents, and the whole system is oriented to provide intuitive and powerful ways of manipulating them in order to achieve a given purpose. This means that processing the usage the researcher makes of a given document can indicate how relevant it is for him/her. To reach this goal, it is essential to track the actions performed by the user when he/she interacts with the system, as done in the history of information-seeking activities explained in Section 5.1.4. Next are provided some examples of how these actions can determine the relevance of a document:

1. If a document is usually opened by a researcher, then it means that he/she needs reading it —to a greater or lesser extent— in order to achieve—at least partly—a given purpose. As the absolute number of times the document is opened cannot provide any information by itself, a more contextualized metric is used: the proportion of information-seeking activities in which the document is opened. The linguistic terms that define the fuzzy input variable —rate of times the document is opened— related to this aspect are:

   - **low rate** is calculated with the function: (0, 1) (0.1, 1) (0.3, 0)
   - **medium rate** is calculated with the function: (0.2, 0) (0.3, 1) (0.4, 1) (0.5, 0)
   - **high rate** is calculated with the function: (0.4, 0) (0.6, 1) (1, 1)

Figure 6.13 illustrates the definition of the variable.
2. Once a document is opened, the researcher can obviously read it, but also take some notes —either by directly annotating it or by associating a note to it in the system— or highlight the parts of the document in order to make them more visible in future readings. Both cases imply that the researcher finds the document relevant enough to spend some time annotating or highlighting it. Additionally, the hypothesis is that, the more the researcher annotates or highlights a document, the more relevant the latter is. In order to define the two input fuzzy variables associated to these aspects, the following definition is proposed:

- **Few times** is calculated with the function: \((0, 1) (1, 1) (2, 0)\)
- **Some times** is calculated with the function: \((1, 0) (2, 1) (5, 1) (6, 0)\)
- **Many times** is calculated with the function: \((5, 0) (7, 1) (\infty, 1)\)

Figures 6.14a and 6.14b illustrate these variables. In the case of page-based documents —articles, theses, books and technical reports—, it could be even more interesting taking into account the number of pages of the document, as it is not
the same writing ten comments in an 8-page article than in a book with hundreds of pages. Then, for example, instead of using the absolute number of notes or highlightings, the average number of them per page could be used. It could also be possible to take a smaller unit of measurement, and consider the density of notes or highlightings per word. This, then, would serve for any document, regardless it is page-based or not.

3. Another action that can reflect that a document is somewhat relevant for the researcher is to get printed. Even if it is difficult to know which is the actual motivation for the researcher to print a document, it seems reasonable to think that the researcher, at least, is going to read it. Additionally, it is highly possible that he/she has printed to work on it, as many researchers, according to the qualitative study, indicated that, if they have to exhaustively read and analyze a document, they prefer to do it in a physical environment. Then, a fuzzy input variable is defined to indicate if a document has been printed or not. Even if in this case it would not be necessary to apply fuzzy logic as the value is binary, and then only two values can occur, the variable has to be fuzzified in order to be usable when calculating the fuzzy output variable "document’s relevance based on the action history of the researcher". The fuzzy input variable "document has been printed" is then defined by only two linguistic terms:

- **printed** is associated to the input value 1 —the numerical value ‘1’ is considered equivalent to the boolean value ‘true’—
- **not printed** is associated to the input value 0 —the numerical value ‘0’ is considered equivalent to the boolean value ‘false’—

Figure 6.15 illustrates the membership functions defining the variable.

![Figure 6.15 Membership functions for the input variable 'document has been printed'](image)

4. On the other side, if a document gets repeatedly discarded, either because it does not match the filtering criteria defined by the researcher, or because he/she
manually clears it from the visualization, then it can be supposed that this document is not relevant for the user. Then if the rate of information-seeking activities in which the document is discarded can be determined, it is possible to use it to calculate the fuzzy input variable "rate of times the document is discarded", whose definition is presented below:

- **low rate** is calculated with the function: \((0, 1) (0.2, 1) (0.4, 0)\)
- **medium rate** is calculated with the function: \((0.3, 0) (0.5, 1) (0.7, 0)\)
- **high rate** is calculated with the function: \((0.6, 0) (0.8, 1) (1, 1)\)

Figure 6.16 illustrates this definition.

Figure 6.16 Membership functions for the input variable "rate of times the document is discarded"

5. Finally, in the case a researcher makes a hard delete of a document —that is he/she not only clears it from the visualization, but also from the storage means—, then the system has to record that the document is not relevant at all for the researcher. As the input fuzzy variable "document has been printed", the variable "document has been deleted from the storage means" requires only two linguistic terms to be defined, as it also allows only the two boolean values:

- **not deleted** is obtained when the input value is 0 —the numerical value ‘0’ is considered equivalent to the boolean value ‘false’—
- **deleted** is obtained when the input value is 1 —the numerical value ‘1’ is considered equivalent to the boolean value ‘true’—

Figure 6.17 illustrates these functions.
And these are the rules used to calculate the output variable "the document has been deleted from the storage system"

- **Rule 1:**
  IF (rate_of_times_opened IS high) AND (number_of_notes IS few) AND (number_of_highlightings IS few)
  THEN doc_by_use_relevance IS medium

- **Rule 2:**
  IF (rate_of_times_opened IS high) AND (number_of_notes IS some) OR (number_of_highlightings IS some)
  THEN doc_by_use_relevance IS high

- **Rule 3:**
  IF (rate_of_times_opened IS high) AND (number_of_notes IS many) OR (number_of_highlightings IS many)
  THEN doc_by_use_relevance IS very_high

- **Rule 4:**
  IF (rate_of_times_opened IS low) AND (number_of_notes IS some) OR (number_of_highlightings IS some)
  THEN doc_by_use_relevance IS medium

- **Rule 5:**
  IF (rate_of_times_opened IS low) AND (number_of_notes IS many) OR (number_of_highlightings IS many)
  THEN doc_by_use_relevance IS high

- **Rule 6:**
  IF (number_of_notes IS many) AND
Adaptive relevance estimation: A Fuzzy Logic approach

(number_of_highlightings IS many)

THEN doc_by_use_relevance IS very_high

• Rule 7:
  IF (doc_is_printed IS printed)

  THEN doc_by_use_relevance IS very_high WITH 0.5

• Rule 8:
  IF (rate_of_times_discarded IS low) AND
    (rate_of_times_opened IS high)

  THEN doc_by_use_relevance IS very_high

• Rule 9:
  IF (rate_of_times_discarded IS low) AND
    ((rate_of_times_opened IS medium)

  THEN doc_by_use_relevance IS high

• Rule 10:
  IF (rate_of_times_discarded IS low) AND
    ((rate_of_times_opened IS low)

  THEN doc_by_use_relevance IS medium

• Rule 11:
  IF (rate_of_times_discarded IS medium) AND
    ((rate_of_times_opened IS high)

  THEN doc_by_use_relevance IS high

• Rule 12:
  IF (rate_of_times_discarded IS medium) AND
    ((rate_of_times_opened IS medium)

  THEN doc_by_use_relevance IS medium

• Rule 13:
  IF (rate_of_times_discarded IS medium) AND
    ((rate_of_times_opened IS low)

  THEN doc_by_use_relevance IS low

• Rule 14:
  IF (rate_of_times_discarded IS high) AND


6.3 Researcher-dependent relevance

((rate_of_times_opened IS high) THEN doc_by_use_relevance IS medium

• Rule 15:
  IF (rate_of_times_discarded IS high) AND
  ((rate_of_times_opened IS NOT high)
  THEN doc_by_use_relevance IS low

• Rule 16:
  IF doc_is_deleted IS deleted
  THEN doc_by_use_relevance IS low

Below, Figure 6.18 illustrates the definition of the fuzzy output variable “document’s relevance based on the action history of the researcher”:

• low relevance is calculated with the function: (0, 1) (0.2, 1) (0.3, 0)
• medium relevance is calculated with the function: (0.2, 0) (0.3, 1) (0.5, 1) (0.6, 0)
• high relevance is calculated with the function: (0.5, 0) (0.6, 1) (0.8, 1) (0.9, 1)
• very_high relevance is calculated with the function: (0.8, 0) (0.9, 1) (1, 1)

Figure 6.18 Membership functions of the output fuzzy variable ’document’s relevance based on the action history of the researcher’

Below are presented some examples of calculation of the output variable “document’s relevance based on the action history of the researcher”. Let’s consider that a researcher has performed 10 information-seeking activities, carrying out the following actions related to a given document:

1. The document is filtered 9 times, either manually or after applying the filtering mechanisms. The only times the document is not filtered, it gets opened by the researcher, who writes a note indicating why the document is not relevant. After that, the researcher deletes the document from the system as he/she believes it is
not going to be relevant even in the future. In this case, the membership levels assigned to each linguistic term by the fuzzy system are:

- low relevance = 1.0
- medium relevance = 0.0
- high relevance = 0.0
- very high relevance = 0.0

The center of gravity of the area displayed in Figure 6.19, 0.1264, represents the defuzzified value of the output variable “document’s relevance based on the action history of the researcher”.

2. The document is filtered by the system 4 times, and 2 more times by the researcher. During the 4 remaining times, the document is opened 2 times, during which the researcher adds 4 notes and highlights 3 parts of the document. In this case, the membership levels assigned to each linguistic term by the fuzzy system are:

- low relevance = 0.1256
- medium relevance = 0.8743
- high relevance = 0.0
- very high relevance = 0.0

In this case, the center of gravity of the resulting membership function —displayed in Figure 6.20— is located at 0.2874.

3. The document is filtered 3 times by the system. Six of the seven remaining times, the researcher opens the document, writes 6 notes and performs 5 highlightings. Finally, he/she prints the document. In this case, the membership levels assigned to each linguistic term by the fuzzy system are:
6.3 Researcher-dependent relevance

- low relevance = 0.0
- medium relevance = 0.0
- high relevance = 1.0
- very high relevance = 0.0

Combining all these values, the final value of the output variable “document’s relevance based on the action history of the researcher” is located at membership level 0.741 of the area illustrated in Figure 6.21.

4. The document is never filtered by the system nor by the researcher. The document is opened 8 times, gets annotated 12 times, and highlighted 7 more times. The researcher also prints the document. In this case, the membership levels assigned to each linguistic term by the fuzzy system are:

- low relevance = 0.0
- medium relevance = 0.0
- high relevance = 0.0
- very high relevance = 1.0

The center of gravity to the area displayed in Figure 6.22, 0.922, represents the final value of the output variable “document’s relevance based on the action history of the researcher”.

This output fuzzy variable is going to be used as input fuzzy variable to calculate the final relevance of documents —see Section 6.5.1. To do so, it is just needed to define the input fuzzy variable just like the output —that is with the same membership functions assigned to the same linguistic terms. In this case, then, the value used as input for this input variable would be the value returned by the fuzzy system after defuzzifying the output variable —that is 0.1264, 0.2874, 0.741, and 0.922 in the examples provided just before.
6.3.2 Document’s relevance based on its thematic similarity with researcher’s topics of interest

As mentioned in Chapters 3 and 5, in order to model a topic and the researcher’s topics of interest, the relevant terms that are associated to it—that is its specific terminology—have to be identified in order to represent them through a vector. As already explained, in this approach—Vector Space Model—each dimension of the vector represents a specific term, and its content represents the frequency of appearance of that term in a given context. In this case, the vectors contain the frequency of occurrence of the terms in the documents addressing the topic being modeled. Then, it is essential to select in the most accurate possible way only those documents that are really representative of the topic and that clearly address it.

As both the researcher’s topics of interest and document’s topic can be represented in the Vector Space Model, their similarity can be calculated by applying the cosine metric, as explained in Section 3.4.3 in Chapter 3. As a result, then, a value between 0 and 1 is obtained, reflecting how similar the document and the researcher’s topics of interest are. This value, then, can be used as input to define a new input fuzzy variable named "document’s thematic similarity with researcher’s topics of interest", whose linguistic terms are listed below and displayed in Figure 6.23:

- **low similarity** is calculated with the function: (0, 0) (0.2, 1) (0.3, 1) (0.4, 0)
- **medium similarity** is calculated with the function: (0.3, 0) (0.4, 1) (0.5, 1) (0.6, 0)
- **high similarity** is calculated with the function: (0.5, 0) (0.6, 1) (0.7, 1) (0.8, 0)
- **very_high similarity** is calculated with the function: (0.7, 0) (0.8, 1) (1, 1)

![Figure 6.23 Membership functions for the input variable "document’s thematic similarity with researcher’s topics of interest"

This variable is going to be one of the input fuzzy variables used to calculate the output fuzzy variable "Document’s relevance" defined in 6.5.1.
6.3.3 Explicit relevance assigned by the researcher

Researchers have to be able to manually indicate how relevant is a document, author, publication venue, or research group. As the system has to adapt to the user’s preferences, this value must prevail over all other inferred or calculated values, like the document’s relevance according to the use the researcher has made of it. A simple way to implement it is allowing the researcher to indicate which of the mentioned elements is a favorite for him/her. However, in our opinion this approach is unrepresentative of reality, as a researcher can want to assign different levels of relevance, ranging from the absolute absence of relevance, to the highest degree of it, which would be equivalent to marking the element as favorite. This degree could be indicated by the researcher through a numerical scale, but, as relevance is an abstract concept, it would probably be difficult for him/her to provide a concrete value indicating the relevance of a document, author, publication venue or research group. In order to address this difficulty, two options are proposed.

First of all, the system can provide the researcher with a more visual approach to indicate the relevance of a concept. As an example, Pazmiño (2016) has proposed to locate visual representations in a 3D virtual environment, and to exclusively use the vertical axis —height— to reflect the relevance of the elements. The relevance of an element, then, can be modified by the user just by "grabbing" its virtual representation and moving it along the vertical axis. With this, an intuitive mapping is proposed, as the higher the visual representation is, the more relevant the element is. This approach can facilitate the assignment of relevance by the researcher, as it is more intuitive and flexible to reflect the relevance of an element by its height than with a numerical value.

In this case, it is only needed to define an input fuzzy variable allowing to determine the "explicit relevance assigned by the researcher" based on the inputted numerical value —that is, the height. If the height value is normalized so that the highest value is represented by the number 1, and lowest one by the number 0, the input fuzzy variable can be defined as it follows:

- **Very high relevance** is represented by the membership function (0,1) (0.25,1) (0.5,0)
- **high relevance** is represented by the membership function (0.25,0) (0.5,1) (0.75,0)
- **medium relevance** is represented by the membership function (0.5,0) (0.75,1) (1,0)
- **low relevance** is represented by the membership function (0.75,0) (1,1)

Figure 6.24 illustrates the membership function of this variable.

Another solution is to allow the researcher indicating the relevance through natural language expressions, like “Very little relevant” or “Very relevant”. It is obviously easier
and more intuitive for a user to indicate that a document is quite relevant, than to define that its relevance is 0.7, for example. This occurs because these natural language expression map better with the human’s mental model than using a simple number. In this case, it is only required to translate each of these terms to the value returning the highest membership level of the corresponding linguistic term:

- the expression *very little relevant* has to be transformed into the value 0.25
- the expression *somewhat relevant* has to be transformed into the value 0.5
- the expression *relevant* has to be transformed into the value 0.75
- the expression *very relevant* has to be transformed into the value 1

In both cases, the variable is going to be one of the input fuzzy variables used to calculate the respective output fuzzy variable — *Document’s relevance, Author’s relevance, Publication Venue’s relevance* or *Research Group’s relevance* — defined in Section 6.5.

### 6.4 Inter-concepts influence

As reflected in Figure 6.25 — which is derived from the document’s model presented in Section 3.3 of Chapter 3 —, the entities that intervene in the information-seeking process, and whose relevance are being calculated in this chapter, are strongly interrelated: a research group is formed by authors, a document addresses a topic and is written by an author, publication venues contain documents, etc. Because of this, besides considering the objective and subjective aspects of the relevance, this dependency has to be taken into account. In fact, it is realistic to state that a document is probably more relevant if it has been written by a relevant author, or that relevant publication venues usually contain more relevant documents.
In this work it is proposed to combine all the aspects that are related in some way to the relevance in order to obtain a more robust result that adapts to the user through the use of a dynamic relevance calculation system able to update the relevance of all the entities that are related. As an example, if a researcher indicates that a document is extremely relevant for him/her, the system should also automatically update, to a certain extent, the relevance of the authors and the publication venue of that document. By doing this, the system would be able to better model the preferences of the researcher, and to better adapt to his/her needs both by giving a special relevance in the visualization of the entities with higher relevance, and by inferring which new entities may be of interest for the researcher..

However, the relationship between concepts is bidirectional, and then there a risk of having infinite mutual propagation of relevance, leading to totally wrong relevance values. Then, inter-concepts dependence has to be controlled in order to avoid, for example, that after increasing the relevance of the documents written by author because the researcher has defined him/her as a favorite, the author’s relevance gets increased again due to the new relevances of the documents he/she has written.

Avoiding the mutual influence between the same two elements solves this problem. This means that if the modification of the relevance of an element A causes the modification of the relevance of an element B, then the latter cannot propagate back its relevance modification to the first one.

In the following section are detailed how the concepts’ relevances are influenced by the other concepts, and how this component of the relevance is integrated with the two other components in order to calculate the final relevance of a document, author, publication venue, or research group.
6.5 Final relevance

6.5.1 Document’s relevance

Based on the dependencies illustrated in Figure 6.25, there are four concepts that can affect the relevance of a document:

1. The amount of times the document has been cited by other documents

The input fuzzy variable defined in Section 6.2.1 is used, but redefining the variable so that the relevance of the citing documents is taken into account. To do so, the weight of the citations coming from relevant documents multiplied, whereas the weight of the citations coming from low relevant documents decreases, even until the influence is equal to 0. The applied formula is:

- IF $\text{citing\_document\_relevance} \geq 0.9$
  THEN $\text{new\_number\_citations} = \text{number\_citations} \times 3$

- IF $0.7 \leq \text{citing\_document\_relevance} < 0.9$
  THEN $\text{new\_number\_citations} = \text{number\_citations} \times 2$

- IF $0.5 \leq \text{citing\_document\_relevance} < 0.7$
  THEN $\text{new\_number\_citations} = \text{number\_citations}$

- IF $0.3 \leq \text{citing\_document\_relevance} < 0.5$
  THEN $\text{new\_number\_citations} = \text{number\_citations} \times 0.5$

- IF $\text{citing\_document\_relevance} < 0.3$
  THEN $\text{new\_number\_citations} = 0$

Nonetheless, using an absolute value as a reference to determine the relevance of the documents citing a giving document may be too simplistic and even lead to confusion. As an example, even if applying the adjustment proposed above, 12 documents with a low relevance —between 0.3 and 0.5— would have the same influence than 2 extremely relevant documents —above 0.9 of relevance. To mitigate this problem, the recalculated variable counting the number of citations weighted according to the relevance of the citing documents is normalized using the following formula:

$$\text{citations\_relevance\_average} = \frac{\text{new\_number\_citations}}{\text{number\_citations}}$$ (6.1)

By doing this, the variable considers both the quantity and relevance of the citations in absolute terms, but also at a more individualistic level as it reflects
the average relevance per citations. The result is that a document that is really
relevant would be very cited, and many of these citations should be also relevant
—the average value, then, would be well above 1. On the contrary, if a document
has been cited by lots of documents that are not relevant, this means that,
probably, the evaluated document is not too relevant —in this case, the average
value would clearly be under 1.

As this new variable is totally different from the previous one, the input fuzzy
variable 'number of citations' has to be redefined. First of all, its name gets
changed to better reflect its content: 'mean relevance citation'. Then, new
linguistic terms with new membership levels are assigned to it, as illustrated in
Figure 6.26:

- low relevance is calculated with the function: \( (0, 1)(0.5, 1)(1, 0) \)
- medium relevance is calculated with the function: \( (0.5, 0)(1, 1)(1.5, 1)(2, 0) \)
- high relevance is calculated with the function: \( (1.5, 0)(2, 1)(2.5, 1)(3, 0) \)
- very_high relevance is calculated with the function: \( (2.5, 0)(3, 1) \)

![mean_relevance_per_citation](image)

Figure 6.26 Membership functions for the input variable 'mean relevance of citation to
document'

2. The relevance of the venue where the document has been published

A document may be considered relevant if it has been published in a relevant
publication venue. Analogously, a publication venue is relevant if its published
documents are good, interesting and relevant enough to be read and referred by
many other documents. Currently, it is quite common to evaluate the quality and
relevance of a document based on the impact metric value of the venue where it
has been published or presented —for example the JCR ranking. If this approach
is followed, as these values are associated to a given range of time —generally
natural years—, it is necessary to take the value of the impact metric in the year
of the publication or presentation of the document.
In this work, the final relevance of a publication venue described in Section 6.5.3 has been used to compute this component of the document’s final relevance.

3. The relevance of the document’s authors
A researcher can consider relevant a document if it has been written by one or more relevant authors. As mentioned in Chapter 5, in the research community there are many authors that are considered not only experts in one or more research topics, but also leaders whose contributions are relevant enough to make the state-of-the-art advance. This makes them have a great credibility resulting in a better acceptance of their works and higher possibility to get their documents referenced by other documents. How to compute the relevance of the authors can be made in different ways:

- The system can only consider the relevance of the main author. This could be valid as the main author is usually the one who the weight of the publication falls on. This approach, however, has a big drawback, as some publication venues do not order the authors according to their weight in the publication, but according to other criteria, like alphabetizing.
- All authors are considered, and their relevances have the same weight on the result.
- All authors are equally considered, except the main author, whose relevance has more influence.
- On occasion, the only presence of a very relevant author, regardless its authorship position and who are the other authors, is used as criterion by a researcher to consider a document, at least, as potentially relevant.

In this work, the second option has been adopted, because it is the most representative and general, as sometimes the first author is not the main one, and then it is not possible to computationally identify him/her —this occurs, for example, when authors are alphabetically sorted. To be more concrete, the final relevance of an author described in Section 6.5.2 has been used.

4. The topic addressed by the document
As already mentioned in Section 6.3.2, the relevance of a document also depends on the topics it addresses: the more these topics are close to the ones in which the researcher is interested in, the more relevant the document is for the researcher.

As a summary, considering the objective and user-dependent relevances, and the ones coming from the related concepts, the relevance of a document can be split into six
components. In order to calculate the document’s relevance based on these components, an input fuzzy variable has to be defined for each of them:

1. The average relevance of the documents citing the evaluated document
   
   • *mean relevance of citation:* defined in this section —see Figure 6.26—

2. The relevance of the document derived from the use the researcher has made of it when performing information-seeking activities in the system
   
   • *document’s relevance based on the action history of the researcher:* defined in Section 6.3.1

3. The degree of similarity between the document’s content and the topics in which the researcher is interested
   
   • *document’s relevance based on its thematic similarity with researcher’s topics of interest:* defined in Section 6.3.2

4. The explicit relevance the user has assigned to the document
   
   • *explicit relevance assigned by the researcher:* defined in Section 6.3.3

5. The relevance of the authors of the document
   
   • *relevance of the document’s authors:* defined in Section 6.5.2

6. The relevance of the venue where the document has been published
   
   • *relevance of the document’s publication venue:* defined in Section 6.5.3

At this point, once all the input fuzzy variables have been defined, a set of rules has to be provided in order to calculate the value of the output fuzzy variable *document’s relevance*. Below are listed the linguistic terms, with their corresponding membership functions, defining the variable, which is illustrated in Figure 6.27.

• *low relevance* is associated to the function: \((0, 1) (0.2, 1) (0.3, 0)\)

• *medium relevance* is associated to the function: \((0.2, 0) (0.3, 1) (0.45, 1) (0.55, 0)\)

• *high relevance* is associated to the function: \((0.45, 0) (0.55, 1) (0.7, 1) (0.8, 0)\)

• *very high relevance* is associated to the function: \((0.7, 0) (0.8, 1) (1, 1)\)
In order to define the rules, two approaches can be followed. On the one hand, it can be assumed that the six components do not affect the relevance calculation in an independent way, but instead they interleave in many ways. This means that the antecedent of a single rule can take into account the value of different components. In fact, this is what has been done to calculate some of the fuzzy variables presented before—"relevance of a document based on the action history of the researcher", "relevance of an author based on his/her impact metrics", and "relevance of a publication venue based on its impact metrics".

However, due to the number of components that intervene in the calculation of the final relevance of a document, this approach implies defining a huge amount of rules to cover all the possible combinations. In fact, considering that each of the six fuzzy variables intervening in the calculation has four possible values, then \(4^6 = 4096\) combinations have to be covered—in Appendix D is listed a very few amount of rules that cover some of these combinations. In fact, even if a rule can cover multiple combinations, there is still a huge amount of rules that have to be defined.

Moreover, there is an important drawback related to the personalization of the relevance calculation. The definition of the fuzzy variables and of the rules used to operate with them has to be predefined by the system’s designer, and then, even if it is tried to maintain the objectivity, the definition is obviously influenced by the designer’s own mental model and preferences. To overcome this problem, the definition of the fuzzy system should also adapt to the researcher’s preferences. To do so, the system should be able to determine not only which aspects are more relevant for the researcher, but also to identify to what extent the researcher wants each of these aspects to influence the calculation of the final relevance and under which conditions—that is, with what rules. Using rules that depend on several components make this approach almost impossible, as the system should be able to infer not only how the user wants to combine these components for each situations, but also to know to which extent he/she want the rule to influence the final relevance calculation.

The second option consists in considering that the final relevance of a document can be cut into several parts, each of them corresponding to one of the components. This
implies that each rule would only make use of one of the fuzzy variables in its antecedent. This solves the two problems stated previously, as in this case only \( 4 \times 6 = 24 \) rules have to be designed —4 values for each of the 6 components—, at maximum, and the system only requires to know from the researcher which weight he/she wants to assigns to each of the six components in order to calculate the final relevance. As in the rules definition it is possible to assign a weight to rule, this second option is chosen as it simplifies the definition of the rules allowing to calculate the "document’s final relevance" and it allows adapting this calculation to the researcher preferences. These, then, are the rules proposed to calculate the relevance of a specific document:

- **Rule 1:**
  
  \[
  \text{IF (mean\_relevance\_of\_citation IS low)}
  \]

  \[
  \text{THEN doc\_relevance IS low WITH } w_1
  \]

- **Rule 2:**

  \[
  \text{IF (mean\_relevance\_of\_citation IS medium)}
  \]

  \[
  \text{THEN doc\_relevance IS medium WITH } w_1
  \]

- **Rule 3:**

  \[
  \text{IF (mean\_relevance\_of\_citation IS high)}
  \]

  \[
  \text{THEN doc\_relevance IS high WITH } w_1
  \]

- **Rule 4:**

  \[
  \text{IF (mean\_relevance\_of\_citation IS very\_high)}
  \]

  \[
  \text{THEN doc\_relevance IS very\_relevant WITH } w_1
  \]

- **Rule 5:**

  \[
  \text{IF (document’s\_relevance\_based\_on\_the\_action\_history\_of\_the\_researcher IS low)}
  \]

  \[
  \text{THEN doc\_relevance IS low WITH } w_2
  \]

- **Rule 6:**

  \[
  \text{IF (document’s\_relevance\_based\_on\_the\_action\_history\_of\_the\_researcher IS medium)}
  \]

  \[
  \text{THEN doc\_relevance IS medium WITH } w_2
  \]

- **Rule 7:**

  \[
  \text{IF (document’s\_relevance\_based\_on\_the\_action\_history\_of\_the\_researcher IS high)}
  \]

  \[
  \text{THEN doc\_relevance IS high WITH } w_2
  \]
• Rule 8:
  IF (document’s relevance based on the action history of the researcher IS high)
  THEN doc_relevance IS high \text{ WITH } w_2

• Rule 9:
  IF (document’s relevance based on its thematic similarity with researcher’s topics of interest IS low)
  THEN doc_relevance IS low \text{ WITH } w_3

• Rule 10:
  IF (document’s relevance based on its thematic similarity with researcher’s topics of interest IS medium)
  THEN doc_relevance IS medium \text{ WITH } w_3

• Rule 11:
  IF (document’s relevance based on its thematic similarity with researcher’s topics of interest IS high)
  THEN doc_relevance IS high \text{ WITH } w_3

• Rule 12:
  IF (document’s relevance based on its thematic similarity with researcher’s topics of interest IS very high)
  THEN doc_relevance IS very high \text{ WITH } w_3

• Rule 13:
  IF (explicit relevance assigned by the researcher IS low)
  THEN doc_relevance IS low \text{ WITH } w_4

• Rule 14:
  IF (explicit relevance assigned by the researcher IS medium)
  THEN doc_relevance IS medium \text{ WITH } w_4

• Rule 15:
  IF (explicit relevance assigned by the researcher IS high)
  THEN doc_relevance IS high \text{ WITH } w_4

• Rule 16:
  IF (explicit relevance assigned by the researcher IS very high)
  THEN doc_relevance IS very high \text{ WITH } w_4
6.5 Final relevance

- Rule 17:
  IF (relevance_of_the_document’s_authors IS low)
  THEN doc_relevance IS low WITH $w_5$

- Rule 18:
  IF (relevance_of_the_document’s_authors IS medium)
  THEN doc_relevance IS medium WITH $w_5$

- Rule 19:
  IF (relevance_of_the_document’s_authors IS high)
  THEN doc_relevance IS high WITH $w_5$

- Rule 20:
  IF (relevance_of_the_document’s_authors IS very_high)
  THEN doc_relevance IS very_high WITH $w_5$

- Rule 21:
  IF (relevance_of_the_document’s_publication_venue IS low)
  THEN doc_relevance IS low WITH $w_6$

- Rule 22:
  IF (relevance_of_the_document’s_publication_venue IS medium)
  THEN doc_relevance IS medium WITH $w_6$

- Rule 23:
  IF (relevance_of_the_document’s_publication_venue IS high)
  THEN doc_relevance IS high WITH $w_6$

- Rule 24:
  IF (relevance_of_the_document’s_publication_venue IS very_high)
  THEN doc_relevance IS very_high WITH $w_6$

In this case, then, the weighting of each is rule is not predefined in the fuzzy system definition, as it depends on the researcher’s preferences. Additionally, as each of the fuzzy variables covers part of the final document’s relevance, the sum of their weights has to be equal to 1.

Then, the system provides two values of the document’s relevance:

- a linguistic value that is easily understandable by the researcher,
- a numerical value that allows the system classifying different documents according to their final relevance.
Next is shown how the relevance of a document would be calculated in different contexts:

1. Let’s imagine that a junior researcher $R_1$ has started to use the information system some months ago. During this time, he/she has accumulated and processed many documents, and he/she feels a little bit overwhelmed, so he/she would like to organize them according to their relevance. Due to his/her low experience, the system considers that all the components determining the relevance of a document should weight the same. Besides, for the same reason, the researcher has not felt confident enough yet to provide an explicit relevance to any document, so this component is not taken into account. Then, the remaining five variables have a weight of 20% in the calculation of the document’s relevance —then, $w_1 = w_2 = w_3 = w_5 = w_6 = 0.2$, and $w_4 = 0$. Let’s evaluate the relevance of three documents in this context:

   (i) The first document $D_1$ has the following characteristics:
   
   - its mean_relevance_of_citation is 35% low and 65% medium;
   - its relevance_based_on_the_action_history_of_the_researcher is 100% low;
   - its relevance_based_on_its_thematic_similarity_with_researcher’s_topics_of_interest is 100% high;
   - the relevance_of_the_document’s_authors is 100% medium;
   - the relevance_of_the_document’s_publication_venue is 69.9% medium and 30.1% high.

   For these values, the fuzzy system returns the result displayed in Figure 6.28.

   ![Figure 6.28 Resulting area of the target output variable "document D_1 relevance" for researcher R_1](image)

   The center of gravity of this area is 0.3953, which corresponds to the following membership levels of the fuzzy output variable "document’s relevance":
6.5 Final relevance

- low relevance: 0.0
- medium relevance: 1.0
- high relevance: 0.0
- very_high relevance: 0.0

Then, it can be concluded that document $D_1$ has a medium relevance for researcher $R_1$.

(ii) The second document $D_2$ has the following characteristics:
- its mean_relevance_of_citation is 80% medium and 20% high;
- its relevance_based_on_the_action_history_of_the_researcher is 87.9% medium and 12.1% low;
- its relevance_based_on_its_thematic_similarity_with_researcher’s_topics_of_interest is 100% very high;
- the relevance_of_the_document’s_authors is 100% low;
- the relevance_of_the_document’s_publication_venue is 100% low.

For these values, the fuzzy system returns the result displayed in Figure 6.29.

![Figure 6.29 Resulting area of the target output variable ’document D2 relevance’ for researcher R1](image)

The center of gravity of this area is **0.4824**, which corresponds to the following membership levels of the fuzzy output variable ’document’s relevance’:
- low relevance: 0.0
- medium relevance: 0.6756
- high relevance: 0.3244
- very_high relevance: 0.0

Then, it can be concluded that document $D_2$ mainly has a medium relevance for researcher $R_1$, but also a slightly high relevance.

(iii) The third document $D_3$ has the following characteristics:
- its mean_relevance_of_citation is 100% medium;
1. Its relevance\_based\_on\_the\_action\_history\_of\_the\_researcher is 100% high;

2. Its relevance\_based\_on\_its\_thematic\_similarity\_with\_researcher’s\_topics\_of\_interest is 50% medium and 50% high;

3. The relevance\_of\_the\_document’s\_authors is 100% very high;

4. The relevance\_of\_the\_document’s\_publication\_venue is 100% low.

For these values, the fuzzy system returns the result displayed in Figure 6.30.

![Figure 6.30 Resulting area of the target output variable "document D₃ relevance" for researcher R₁](image)

The center of gravity of this area is 0.5, which corresponds to the following membership levels of the fuzzy output variable "document’s relevance":

- low relevance: 0.0
- medium relevance: 0.5
- high relevance: 0.5
- very\_high relevance: 0.0

Then, it can be concluded that document D₃ has a medium-high relevance for researcher R₁.

2. Let’s imagine a researcher R₂ that has been using the information system for a long time. The researcher knows the system has properly identified his/her topics of interest during this time. As he/she wants to discover new relevant documents related to his/her topics of interest, he/she indicates the system that the similarity with these topics has to be considerably taken into account. Besides, he/she also would like to get documents whose authors are relevant and that have been widely cited by other relevant documents. Finally, he/she also wants the system to take into consideration the publication venue, but at a low level. The history of use and explicit relevance are not taken into account, as he/she looks for new documents, and then they have supposedly not been processed by him/her yet.
According to these restrictions, the following weights can be assigned: \( w_1 = 0.25, \ w_3 = 0.4, \ w_5 = 0.25, \ w_6 = 0.1, \) and \( w_2 = w_4 = 0. \) Let’s evaluate the relevance of three documents in this context:

(i) For document \( D_1 \) — whose characteristics have been described in the previous example of researcher \( R_1 \) —, the fuzzy system determines that its relevance for researcher \( R_2 \) is the one displayed in Figure 6.31.

![Figure 6.31 Resulting area of the target output variable "document \( D_1 \) relevance" for researcher \( R_2 \)](image)

The center of gravity of this area is \( 0.4931 \), which corresponds to the following membership levels of the fuzzy output variable "document’s relevance":

- low relevance: 0.0
- medium relevance: 0.5691
- high relevance: 0.4309
- very_high relevance: 0.0

Then, it can be concluded that document \( D_1 \) has a medium-high relevance for researcher \( R_2 \).

(ii) For document \( D_2 \) and researcher \( R_2 \), the fuzzy system returns the result displayed in Figure 6.32.

![Figure 6.32 Resulting area of the target output variable "document \( D_2 \) relevance" for researcher \( R_2 \)](image)
Adaptive relevance estimation: A Fuzzy Logic approach

The center of gravity of this area is 0.5471, which corresponds to the following membership levels of the fuzzy output variable "document’s relevance":

- low relevance: 0.0
- medium relevance: 0.02857
- high relevance: 0.9714
- very_high relevance: 0.0

Then, it can be concluded that document $D_2$ has a high relevance for researcher $R_2$.

(iii) Finally, for document $D_3$, researcher $R_2$ gets the result displayed in Figure 6.33 from the fuzzy system.

![Figure 6.33 Resulting area of the target output variable "document $D_3$ relevance" for researcher $R_2$](image)

The center of gravity of this area is 0.5555, which corresponds to the following membership levels of the fuzzy output variable "document’s relevance":

- low relevance: 0.0
- medium relevance: 0.0
- high relevance: 1.0
- very_high relevance: 0.0

Then, it can be concluded with total certainty that document $D_3$ has a high relevance for researcher $R_2$.

3. The last researcher, $R_3$, is a senior researcher that has input all the documents of his/her local collection into the information system. As he/she perfectly knows which are the most relevant documents in his/her field of researcher, he/she assigned explicit relevance to them, and he/she wants the system to especially take this aspect into consideration to calculate the final relevance of the documents. Because of this, he/she does not want to provide to much weight on the relevance of the citing documents. As he/she has recently inputted the documents, most of the use he/she has made of them has been done before, and then the system has
not recorded how he/she has used them. As all documents belong to his/her local collection, he/she knows the all are somehow related to his/her topics of interests, so he/she does not want to give too much importance to this aspect. Finally, he/she wants author’s relevance to have more impact than publication venue’s relevance. That’s why this component is not taken into account. According to these restrictions, the following weights can be assigned: $w_1 = w_3 = w_6 = 0.1$, $w_4 = 0.5$, $w_5 = 0.2$, and $w_2 = 0$. Let’s evaluate the relevance of three documents in this context:

(i) Using the same document $D_1$, to which researcher $R_3$ has assigned a medium explicit relevance. The fuzzy system returns the result displayed in Figure 6.34.

![Figure 6.34 Membership functions for the target output variable 'document $D_1$ relevance' for researcher $R_3$](image)

The center of gravity of this area is $0.4055$, which corresponds to the following membership levels of the fuzzy output variable 'document’s relevance’:

- low relevance: 0.0
- medium relevance: 1.0
- high relevance: 0.0
- very_high relevance: 0.0

As a result, document $D_1$ has a medium relevance for researcher $R_3$, which coincides with the explicit relevance he/she has assigned to the document.

(ii) In this case, researcher $R_3$ assigns a high relevance to document $D_2$, and the fuzzy system returns the result displayed in Figure 6.35.
The center of gravity of this area is 0.5197, which corresponds to the following membership levels of the fuzzy output variable "document's relevance":

- low relevance: 0.0
- medium relevance: 0.3026
- high relevance: 0.6974
- very_high relevance: 0.0

Then, it can be concluded that, to a large degree, the fuzzy system confirms the explicit relevance assigned by researcher $R_3$ to document $D_2$, as it rates it as highly relevant. However, it also determines that the relevance document is of medium with 30%. This mainly occurs because the relevance of the authors and of the publication venue is low, which makes the whole relevance decrease despite the opinion of researcher $R_3$.

(iii) The third document $D_3$ is evaluated as very relevant by researcher $R_3$, and gets as result result the membership function displayed in Figure 6.36.

The center of gravity of this area is 0.7591, which corresponds to the following membership levels of the fuzzy output variable "document's relevance":

- low relevance: 0.0
- medium relevance: 0.0
Then, it can be concluded that the fuzzy system mainly assigns to document $D_3$ the same relevance researcher $R_3$ has assigned to it —very high relevance—, but it also considers that the document has a high relevance, with around 41% of certainty, because of the low and medium relevances of the publication venue and of the citing documents.

6.5.2 Author’s relevance

Based on the dependencies illustrated in Figure 6.25, there are two aspects that can affect the relevance of an author:

1. The amount of the documents the author has written and their relevance
   As, ultimately, a document reflects the hypothesis and proposals of one or more authors, an author may be considered relevant if many of the documents he/she has written are in turn relevant. Then, for example, an author whose documents are cited many times can be considered, in general, a relevant author.

As for the calculation of the document’s relevance depending on the amount of times it has been cited, and on the relevance of the citing documents themselves, in this case the relevances of all the documents written by the author have also to be somehow integrated. Actually, most of the impact metrics used in the research field to indicate the relevance of an author —including those presented in the Section 6.2.3— are in fact based on the amount of times his/her research documents have been cited. However, as explained in Section 6.5.1, document’s relevance is not only based on the number of citations, and then the whole relevance of a document, with all its components, can be more representative and illustrative of its influence on author’s relevance calculation.

As set forth in Section 6.2.2, the intermediate step to integrate all the documents’ relevance requires taking into account the individual relevance of each of the documents written by an author in order to recalculate his/her relevance:

- IF $written\_document\_relevance \geq 0.9$
  THEN $new\_amount\_of\_documents = amount\_of\_documents \times 3$
- IF $0.7 \leq written\_document\_relevance < 0.9$
  THEN $new\_amount\_of\_documents = amount\_of\_documents \times 2$
- IF $0.5 \leq written\_document\_relevance < 0.7$
  THEN $new\_amount\_of\_documents = amount\_of\_documents$

• $high\ relevance: 0.4086$
• $very\_high\ relevance: 0.5914$
• IF $0.3 \leq \text{written\_document\_relevance} < 0.5$
  \[
  \text{THEN new\_amount\_of\_documents} = \text{amount\_of\_documents} \times 0.5
  \]

• IF $\text{written\_document\_relevance} < 0.3$
  \[
  \text{THEN new\_amount\_of\_documents} = 0
  \]

As it happens with the document’s input fuzzy variable "mean relevance of citation", using an absolute number, even if it considers the relevance of the written documents, may not properly reflect what it is expected. Then, the same normalization used for "mean relevance of citation" is applied for this variable, consisting in dividing the recalculated new amount of written documents by the real number of written documents. The resulting input fuzzy variable, "mean relevance of written document", is defined as it follows —see Figure 6.37:

• low relevance is calculated with function: (0, 1) (0.5, 1) (1, 0)
• medium relevance is calculated with function: (0.5, 0) (1, 1) (1.5, 1) (2, 0)
• high relevance is calculated with function: (1.5, 0) (2, 1) (2.5, 1) (3, 0)
• very high relevance is calculated with function: (2.5, 0) (3, 1)

![Figure 6.37 Membership functions for the input variable 'mean relevance of documents written by the author'](image)

2. Relevance of the research group to which the author is affiliated

Authors usually are affiliated to research groups, and then the relevance of a research group has to be taken into account to calculate the relevance of its affiliated authors. In this case there are also two approaches to calculate this variable:

• using the relevance of the research group to which the author is currently affiliated, or
• taking into consideration all the research groups to which the author has belonged to. In this case, only the most relevant one can be considered,
the average relevance can be calculated, or even a weighting scheme can be applied so that the influence of the relevance of past research groups is lower.

The first option is the one used in this solution as the current affiliation of an author is considered the most illustrative of the current situation of the author. In this work, the final relevance of a research group described in Section 6.5.4 has been used to compute this component of the author’s final relevance.

Then, considering the objective and user-dependent relevances, and the ones deriving from the related concepts, the relevance of an author can be split into four components. In order to calculate the author’s relevance based on these components, an input fuzzy variable has to be defined for each of them:

1. The average relevance of the documents written by the evaluated author
   - *mean relevance of written document*: defined in this section —see Figure 6.37—

2. The relevance of the author based on his/her impact metrics
   - *author’s relevance based on his/her impact metrics*: defined in Section 6.2.3

3. The author’s relevance based on the relevance of the research group to which he/she is affiliated
   - *relevance of the research group to which the author is affiliated*: defined in Section 6.5.4

4. The explicit relevance the user has assigned to the author
   - *explicit relevance assigned by the researcher*: defined in Section 6.3.3

At this point, once all the input fuzzy variables have been defined, a set of rules have also to be described in order to calculate the value of the output fuzzy variable *author’s relevance*. Below are listed the linguistic terms, with their corresponding membership functions, defining the variable, which is illustrated in Figure 6.38:

- *low relevance* is associated to the function: \((0, 1) (0.2, 1) (0.3, 0)\)
- *medium relevance* is associated to the function: \((0.2, 0) (0.3, 1) (0.45, 1) (0.55, 0)\)
- *high relevance* is associated to the function: \((0.45, 0) (0.55, 1) (0.7, 1) (0.8, 0)\)
- *very_high relevance* is associated to the function: \((0.7, 0) (0.8, 1) (1, 1)\)
As for the document, each of the components of the "author’s relevance" is individually considered, and then only one variable is used in the antecedent of the rules. However, in Appendix D is provided a set of rules that cover some of the $4^4 = 256$ combinations that can be done with the input fuzzy variables. Next are presented the rules proposed to calculate the relevance of a specific author:

- **Rule 1:**
  IF (relevance_by_impact IS low) 
  THEN auth_relevance IS low WITH $w_1$

- **Rule 2:**
  IF (relevance_by_impact IS medium) 
  THEN auth_relevance IS medium WITH $w_1$

- **Rule 3:**
  IF (relevance_by_impact IS high) 
  THEN auth_relevance IS high WITH $w_1$

- **Rule 4:**
  IF (relevance_by_impact IS very_high) 
  THEN auth_relevance IS very_high WITH $w_1$

- **Rule 5:**
  IF (mean_relevance_per_written_document IS low) 
  THEN auth_relevance IS low WITH $w_2$

- **Rule 6:**
  IF (mean_relevance_per_written_document IS medium) 
  THEN auth_relevance IS medium WITH $w_2$

- **Rule 7:**
  IF (mean_relevance_per_written_document IS high) 
  THEN auth_relevance IS high WITH $w_2$
THEN auth_relevance IS high WITH $w_2$

• Rule 8:
  IF (mean_relevance_per_written_document IS very_high)
  THEN auth_relevance IS very_high WITH $w_2$

• Rule 9:
  IF (relevance_by_research_group IS low)
  THEN auth_relevance IS low WITH $w_3$

• Rule 10:
  IF (relevance_by_research_group IS medium)
  THEN auth_relevance IS medium WITH $w_3$

• Rule 11:
  IF (relevance_by_research_group IS high)
  THEN auth_relevance IS high WITH $w_3$

• Rule 12:
  IF (relevance_by_research_group IS very_high)
  THEN auth_relevance IS very_high WITH $w_3$

• Rule 13:
  IF (relevance_by_user IS low)
  THEN auth_relevance IS low WITH $w_4$

• Rule 14:
  IF (relevance_by_user IS medium)
  THEN auth_relevance IS medium WITH $w_4$

• Rule 15:
  IF (relevance_by_user IS high)
  THEN auth_relevance IS high WITH $w_4$

• Rule 16:
  IF (relevance_by_user IS very_high)
  THEN auth_relevance IS very_high WITH $w_4$

As done in the calculation of the document’s relevance in Section 6.5.1, the weighting of each component is not predefined and depends on the researcher’s preferences and purpose.
6.5.3 Publication venue’s relevance

Based on the dependencies illustrated in Figure 6.25, there is only one concept that affects the relevance of a publication venue:

1. Publication venue’s relevance based on the amount of documents it has published and their relevance

Even if the metrics that are proposed to evaluate the relevance of a publication venue already take into account, one way or another, the amount of times the documents published in it have been cited, the full relevance of the document, as defined in this work, has not been considered to evaluate the relevance of a publication venue. That’s why a new input fuzzy variable similar to the author’s relevance based on the amount of the documents he/she has written and their relevance is defined, but in this case considering the documents being published in a given publication venue. As explained in Section 6.2.5, the amount of documents published in a given publication venue has to be recalculated based on their relevance in order to better reflect the actual relevance of the venue:

- IF \( \text{published\_document\_relevance} \geq 0.9 \)
  THEN \( \text{new\_amount\_of\_documents} = \text{amount\_of\_documents} \times 3 \)

- IF \( 0.7 \leq \text{published\_document\_relevance} < 0.9 \)
  THEN \( \text{new\_amount\_of\_documents} = \text{amount\_of\_documents} \times 2 \)

- IF \( 0.5 \leq \text{published\_document\_relevance} < 0.7 \)
  THEN \( \text{new\_amount\_of\_documents} = \text{amount\_of\_documents} \)

- IF \( 0.3 \leq \text{published\_document\_relevance} < 0.5 \)
  THEN \( \text{new\_amount\_of\_documents} = \text{amount\_of\_documents} \times 0.5 \)

- IF \( \text{published\_document\_relevance} < 0.3 \)
  THEN \( \text{new\_amount\_of\_documents} = 0 \)

Again in this case, the previous variable is normalized to provide a more realistic value. Then, a new input fuzzy variable, ’mean relevance of published document’, is defined as it follows —see Figure 6.39:

- low relevance is calculated with the function: \( (0, 1) (0.5, 1) (1, 0) \)
medium relevance is calculated with the function: (0.5, 0) (1, 1) (1.5, 1) (2, 0)

high relevance is calculated with the function: (1.5, 0) (2, 1) (2.5, 1) (3, 0)

very_high relevance is calculated with the function: (2.5, 0) (3, 1)

Figure 6.39 Membership functions for the input variable "mean relevance of documents published by the venue"

Then, considering the objective and user-dependent relevances, and the one coming from the related concepts, the relevance of a publication venue can be split into three components. In order to calculate the publication venue’s relevance based on these components, an input fuzzy variable has to be defined for each of them:

1. The average relevance of the documents published in the venue.
   - mean relevance of published document: defined in this section —see Figure 6.39—

2. The relevance of the publication venue based on his/her impact metrics
   - publication venue’s relevance based on its impact metrics: defined in Section 6.2.4

3. The explicit relevance the user has assigned to the publication venue
   - explicit relevance assigned by the researcher: defined in Section 6.3.3

At this point, once all the input fuzzy variables have been described, a set of rules has to be defined in order to calculate the value of the output fuzzy variable "publication venue’s relevance". Below are listed the linguistic terms, with their corresponding membership functions, defining the variable, which is illustrated in Figure 6.40:

- low relevance is associated to the function: (0, 1) (0.2, 1) (0.3, 0)
• medium relevance is associated to the function: (0.2, 0) (0.3, 1) (0.45, 1) (0.55, 0)
• high relevance is associated to the function: (0.45, 0) (0.55, 1) (0.7, 1) (0.8, 0)
• very_high relevance is associated to the function: (0.7, 0) (0.8, 1) (1, 1)

Figure 6.40 Membership functions for the target output variable "publication venue’s relevance"

As for the document and the author, each of the components of the "publication venue’s relevance" has been considered individually, and then only one variable is used in the antecedent of the rules. However, in Appendix D is presented a list of rules that cover some of the \(4^3 = 64\) combinations that can be done with the input fuzzy variables. Next are presented the rules proposed to calculate the relevance of a specific publication venue:

• Rule 1:
  IF (relevance_by_impact IS low)
  THEN publven_relevance IS low WITH \(w_1\)

• Rule 2:
  IF (relevance_by_impact IS medium)
  THEN publven_relevance IS medium WITH \(w_1\)

• Rule 3:
  IF (relevance_by_impact IS high)
  THEN publven_relevance IS high WITH \(w_1\)

• Rule 4:
  IF (relevance_by_impact IS very_high)
  THEN publven_relevance IS very_high WITH \(w_1\)

• Rule 5:
  IF (mean_relevance_per_published_document IS low)
6.5 Final relevance

THEN publven_relevance IS low WITH $w_2$

- Rule 6:
  IF (mean_relevance_per_published_document IS medium)
  THEN publven_relevance IS medium WITH $w_2$

- Rule 7:
  IF (mean_relevance_per_published_document IS high)
  THEN publven_relevance IS high WITH $w_2$

- Rule 8:
  IF (mean_relevance_per_published_document IS very_high)
  THEN publven_relevance IS very_high WITH $w_2$

- Rule 9:
  IF (relevance_by_user IS low)
  THEN publven_relevance IS low WITH $w_3$

- Rule 10:
  IF (relevance_by_user IS medium)
  THEN publven_relevance IS medium WITH $w_3$

- Rule 11:
  IF (relevance_by_user IS high)
  THEN publven_relevance IS high WITH $w_3$

- Rule 12:
  IF (relevance_by_user IS very_high)
  THEN publven_relevance IS very_high WITH $w_3$

Also in this case, the rules are not weighted in advance in order to allow adapting their influence on the "publication venue’s relevance" calculation according to the researcher’s preferences and purposes.

6.5.4 Research group’s relevance

Based on the dependencies illustrated in Figure 6.25, there is only one concept that affects the relevance of a research group:

1. The amount of authors affiliated to it and their relevance

   As explained in Section 6.2.6, a research group is nothing but a conglomeration of authors that investigate in similar research topics. Then, it seems obvious
that the relevance of a research group is directly dependent on the relevance of the authors that form part of it. However, authors are not always affiliated to the same research group indefinitely. Then, only the relevance achieved by the authors while belonging to the group has to be considered. As in other cases, there are different considerations to take before calculating the relevance of a research group based on its members:

- **Considered period**
  - The variable can take into account the relevance of the group since its foundation, which implies making use of the relevance of all the authors that have made part of the group at some moment.
  - The variable can only make references to the relevance of the authors that are currently affiliated to it.

- **Integration**
  - The relevance of the research group can be the one of the most relevant author.
  - An average relevance can be calculated, either uniformly or weighting their relevance according to the antiquity of the author’s affiliation. It also can be distinguished, in case the whole group’s existence is considered, if the author is still member of the group, or if he/she has quit it.

As for the publication venues, it is assumed that for a researcher it is interesting knowing the current situation of a research group, and then the amount of authors that currently are affiliated to the research group is weighted according to their relevance:

- IF $\text{author\_relevance} \geq 0.9$
  THEN $\text{new\_amount\_of\_authors} = \text{amount\_of\_documents} \times 3$

- IF $0.7 \leq \text{author\_relevance} < 0.9$
  THEN $\text{new\_amount\_of\_authors} = \text{amount\_of\_documents} \times 2$

- IF $0.5 \leq \text{author\_relevance} < 0.7$
  THEN $\text{new\_amount\_of\_authors} = \text{amount\_of\_documents}$

- IF $0.3 \leq \text{author\_relevance} < 0.5$
  THEN $\text{new\_amount\_of\_authors} = \text{amount\_of\_documents} \times 0.5$

- IF $\text{author\_relevance} < 0.3$
  THEN $\text{new\_amount\_of\_authors} = 0$
This variable is normalized in the same manner as done with "mean relevance citation", "mean relevance of written document", and "mean relevance of published document", resulting in the variable "mean relevance of authors affiliated to the research group". Its definition, illustrated in Figure 6.41, contains the following linguistics terms:

- **low relevance** is calculated with function: (0, 1) (0.5, 1) (1, 0)
- **medium relevance** is calculated with function: (0.5, 0) (1, 1) (1.5, 1) (2, 0)
- **high relevance** is calculated with function: (1.5, 0) (2, 1) (2.5, 1) (3, 0)
- **very high relevance** is calculated with function: (2.5, 0) (3, 1)

![Figure 6.41 Membership functions for the input variable 'research group's relevance'](image)

Then, considering both the user-dependent relevance and the inter-concepts dependency, the relevance of a research group can be split into two components. In order to calculate the research group relevance based on these components, an input fuzzy variable has to be defined for each of them:

1. The average relevance of the authors affiliated to the research group
   - **mean relevance of authors affiliated to the research group**: defined in this section —see Figure 6.41—

2. The explicit relevance the user has assigned to the research group
   - **explicit relevance assigned by the researcher**: defined in Section 6.3.3

At this point, once both input fuzzy variables have been defined, a set of rules has to be proposed in order to calculate the value of the output fuzzy variable "research group's relevance". Below are listed the linguistic terms, with their corresponding membership functions, defining the variable, which is illustrated in Figure 6.42:

- **low relevance** is associated to the function: (0, 1) (0.2, 1) (0.3, 0)
medium relevance is associated to the function: (0.2, 0) (0.3, 1) (0.45, 1) (0.55, 0)

high relevance is associated to the function: (0.45, 0) (0.55, 1) (0.7, 1) (0.8, 0)

very_high relevance is associated to the function: (0.7, 0) (0.8, 1) (1, 1)

Figure 6.42 Membership functions for the target output variable "research group’s relevance"

In this case, there are only two input fuzzy variables that intervene in the "research group’s relevance" calculation, which means there are only $4^2 = 16$ combinations of their values. However, as one of the variables is the "explicit relevance assigned by the researcher", and that this variable is consistent enough by itself to define a context, there actually are only two independent variables. This results in the same approach used in the other cases presented below. Next are presented the rules proposed to calculate the relevance of a specific research group:

And these are the rules proposed to calculate the relevance of a specific research group:

- **Rule 1:**
  \[
  \text{IF (mean_relevance_per_affiliated_author IS low)}
  \]
  \[
  \text{THEN resgroup_relevance IS low WITH } w_1
  \]

- **Rule 2:**
  \[
  \text{IF (mean_relevance_per_affiliated_author IS medium)}
  \]
  \[
  \text{THEN resgroup_relevance IS medium WITH } w_1
  \]

- **Rule 3:**
  \[
  \text{IF (mean_relevance_per_affiliated_author IS high)}
  \]
  \[
  \text{THEN resgroup_relevance IS high WITH } w_1
  \]

- **Rule 4:**
  \[
  \text{IF (mean_relevance_per_affiliated_author IS very_high)}
  \]
  \[
  \text{THEN resgroup_relevance IS very_high WITH } w_1
  \]
6.6 Time-dependence of relevance

- Rule 5:
  \[\text{IF (relevance\_by\_user IS low)}\]
  \[\text{THEN resgroup\_relevance IS low WITH } w_2\]

- Rule 6:
  \[\text{IF (relevance\_by\_user IS medium)}\]
  \[\text{THEN resgroup\_relevance IS medium WITH } w_2\]

- Rule 7:
  \[\text{IF (relevance\_by\_user IS high)}\]
  \[\text{THEN resgroup\_relevance IS high WITH } w_2\]

- Rule 8:
  \[\text{IF (relevance\_by\_user IS very\_high)}\]
  \[\text{THEN resgroup\_relevance IS very\_high WITH } w_2\]

As in the other cases, the weighting of the components in the "research group’s relevance" calculation is customizable in order to adapt to the researcher’s preferences and to his/her specific purpose.

Table 6.1 summarizes the components used to calculate the relevance of each type of entity.

6.6 Time-dependence of relevance

In research, the relevance of a document, author, publication venue, or research group is usually linked to time. This happens because of the inner nature and purpose of research, which is discovering and generating new knowledge to make advance the state-of-the-art. Obviously, this does not mean that old contributions are not valuable or needed. On the contrary, some of these contributions have a timeless relevance and are used as the basis allowing researcher starting from fully validated scenarios and trying to complete them.

However, there may be many situations in which a document, author, publication venue, or research group is very relevant, but only for a period of time. As an example, a document may be relevant because the topic it addresses is very innovative and there are not so many publications related to it at that moment; an author may have been very relevant because he/she has written a document that is considered a reference in a topic but it is the only one; or a journal may have been the most relevant venue where most of the published papers have been considered very relevant at a given moment.
Table 6.1 Summary of the components used to calculate the relevance of the different entities implied in research-oriented information-seeking activities

<table>
<thead>
<tr>
<th>Component</th>
<th>Objective relevance</th>
<th>Subjective relevance</th>
<th>Inter-concepts relevance</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>DOCUMENT</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Amount of times the document has been cited</td>
<td>✓</td>
<td></td>
<td>Weighted based on the relevance of the citing documents</td>
</tr>
<tr>
<td>Thematic similarity with researcher’s topics of interest</td>
<td>✓</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Usage history of the document by the researcher</td>
<td>✓</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Explicit relevance assigned by the researcher</td>
<td>✓</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Relevance of the venue where the document has been published</td>
<td></td>
<td>✓</td>
<td></td>
</tr>
<tr>
<td>Relevance of the document’s authors</td>
<td></td>
<td>✓</td>
<td></td>
</tr>
<tr>
<td><strong>AUTHOR</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Amount of documents the author has written</td>
<td>✓</td>
<td></td>
<td>Weighted based on the relevance of the documents</td>
</tr>
<tr>
<td>Impact metrics of the author</td>
<td></td>
<td>✓</td>
<td></td>
</tr>
<tr>
<td>Explicit relevance assigned by the researcher</td>
<td></td>
<td>✓</td>
<td></td>
</tr>
<tr>
<td>Relevance of the research group to which the author is affiliated to</td>
<td></td>
<td></td>
<td>✓</td>
</tr>
<tr>
<td><strong>P. VENUE</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Amount of documents published in the venue</td>
<td>✓</td>
<td></td>
<td>Weighted based on the relevance of the documents</td>
</tr>
<tr>
<td>Impact metrics of the publication venue</td>
<td>✓</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Explicit relevance assigned by the researcher</td>
<td>✓</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>RES. GROUP</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Amount of authors affiliated to the research group</td>
<td>✓</td>
<td></td>
<td>Weighted based on the relevance of the authors</td>
</tr>
<tr>
<td>Explicit relevance assigned by the researcher</td>
<td>✓</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
but that, after the appearance of new journals, has lost the leadership and now accepts papers with a lower relevance.

That’s why, in order to take into account this possibility, in this solution it is proposed to systematically mitigate the relevance of these elements as time passes. By doing this, the elements that are usually used by the researcher —either directly, or by the propagation of the relevance explained before— maintain or increase their relevance, while those that are punctually used or that have been used long time ago can be considered as obsolete and then have to gradually lose their relevance until becoming totally irrelevant.

There is an exception that should be implemented to ensure that the relevance calculated by the system actually fits the one perceived by the researcher. The relevance explicitly assigned by a researcher to a concept does not have to be affected by this mitigation, or if so, the mitigation has to be limited. This is because it reflects a conscious and direct feedback from the researcher, and then there is no doubt about its validity, regardless the moment it has been assigned. This exception aligns with the approach used when calculating the relevance of any of the elements, as the subjective relevance assigned by the researcher prevails over any other inferred or calculated component of the relevance, that in this case would only serve as nuance.

In any case, the system should save both the mitigated and the non-mitigated relevances, as this would allow to analyze how the interests of the researcher have evolved over time, allowing to better determine the relevance of a concept. As an example, let’s suppose that a researcher always cites a given document in the articles he/she writes as, even if it is a very old document that adds nothing new, it is a basic reference that must be cited for completeness. In this case, the user does not make too much use of it apart from citing it, and then the temporal mitigation could result in highly diminishing its relevance, even if it is one of the pillars of the research topic. However, by analyzing its evolution of use, the system could infer that the document has to be present to achieve some purposes, like elaborating a state-of-the-art of the research field.
Chapter 7

Practical applications

The proposed models can be of great utility in different contexts. Next are explained the main ones, together with some illustrative examples.

7.1 Use for formal characterization of information systems

Due to the completeness of the models, they can first of all be used as a framework for characterizing existing information systems addressing information-seeking and/or information visualization for research purposes, by providing a set of concrete, well-defined, non-ambiguous and specific terminology related to the process. For example, the model can be used to describe which information-seeking tasks allows to perform Google Scholar, and which are its main characteristics —see Figure 7.1.

As it can be seen, Google Scholar allows filtering, exploring, reading, storing and managing documents. In the case of the filtering task, for example, 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 has to 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 be included in the resulting document collection.
Figure 7.1 Description of the information-seeking tasks supported by Google Scholar using the conceptual model.
7.2 Use for information systems comparison

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 set of documents iteratively and incrementally, interleaving diverse tasks if needed. Of course, Google Scholar also allows performing a single task without any further refinement.

Google Scholar also allows reading the documents belonging to the filtering result collection, provided they are included in the focus of exploration at that moment. Nonetheless, Google Scholar does not allow to highlight relevant parts of a document, nor adding personal notes to them.

Finally, the user can also archive the the links pointing to the documents he/she considers relevant in a remote means. He/she can also assign a label to these archived documents in order to classify them.

7.2 Use for information systems comparison

As an extension of the previous use, the proposed conceptual models can be used as a thesaurus of concepts intervening in the seeking and/or visualization of information. This is especially interesting to perform comparisons of different competing systems for information-seeking activities in research. As an example, Tables 7.1 and 7.2 comprehensively describe two information visualization systems —Google Scholar and Calimaco, respectively—, which greatly helps the reader to compare in a simple and structured manner their characteristics by identifying their similarities and differences. This can be especially useful to evaluate, among several candidates, which information system may be more suitable to meet all the user’s requirements.

7.3 Use for designing an information system

But, besides using the models as a descriptive tool, they can be used for design purposes. This is especially useful when the information system is been designed following a user-centered design process (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, in order to faithfully provide a solution that covers the actual needs of the users. A conceptual model allows to formally represent the analysis of the information obtained during the data collection phase —in this case, the qualitative study. Besides, one of the most intrinsic characteristics of the
<table>
<thead>
<tr>
<th>Source</th>
<th>Conceptual Framework</th>
<th>Document Representation</th>
<th>Visualized Diagram</th>
</tr>
</thead>
<tbody>
<tr>
<td>Google Scholar</td>
<td>Use of the model as framework to describe Google Scholar</td>
<td>Diagram illustrating the model's application</td>
<td>Model diagram showing key components</td>
</tr>
</tbody>
</table>

Table 7.1: Use of the model as framework to describe Google Scholar
### 7.3 Use for designing an information system

| WHAT ARE YOU TRYING TO \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \*
<table>
<thead>
<tr>
<th>Concept</th>
<th>Relationships/Sub-concepts</th>
<th>Instances</th>
</tr>
</thead>
<tbody>
<tr>
<td>Calimaco</td>
<td>Use of the model as framework to describe Calimaco</td>
<td></td>
</tr>
</tbody>
</table>

Table 7.2 Use of the model as framework to describe Calimaco
| 7.3 Use for designing an information system |

<table>
<thead>
<tr>
<th>Information Visualization</th>
<th>Table</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>E6</strong></td>
<td>E7</td>
</tr>
<tr>
<td><strong>DSS NOT MAAY</strong></td>
<td>Course</td>
</tr>
<tr>
<td><strong>DSS NOT MAAY</strong></td>
<td>Course</td>
</tr>
<tr>
<td><strong>DSS NOT MAAY</strong></td>
<td>Course</td>
</tr>
<tr>
<td><strong>DSS NOT MAAY</strong></td>
<td>Course</td>
</tr>
<tr>
<td><strong>DSS NOT MAAY</strong></td>
<td>Course</td>
</tr>
<tr>
<td><strong>DSS NOT MAAY</strong></td>
<td>Course</td>
</tr>
</tbody>
</table>

**Legend for Table:**

- **DSS NOT MAAY:** Denotes the absence or lack of an item.
- **Course:** Represents a course or module.
- **Information Visualization:** Refers to the process of converting data into graphical representations.

The table outlines various courses and their attributes in the context of information visualization.

---

More detailed explanations and descriptions are available in the accompanying text.