Negotiating a Multidimensional Framework for Relevance Space

Silvia Gabrielli
Department of Psychology
University of Padova
Via Venezia, 8
I 35131 Padova, Italy
ph.: +39 049 8276903, fax: +39 049 8276600
E-mail: gabriell@psico.unipd.it

Stefano Mizzaro
Department of Mathematics and Computer Science
University of Udine
Via delle Scienze, 206 — Loc. Rizzi
I 33100 Udine, Italy
ph.: +39 0432 558456, fax: +39 0432 558499
E-mail: mizzaro@dimi.uniud.it
WWW: http://www.dimi.uniud.it/˜mizzaro

Abstract

This work reports the results of an enquiry on the concept of relevance and on relevance judgments carried out during the MIRA workshops activities in 1998/1999.

Starting from a previous proposal [23], we present the multidimensional relevance space, a framework for describing the various kinds of relevance, which has been negotiated with experts belonging to the MIRA community. The relevance dimensions of information needs, information resources, and information use context are presented, and a three dimensional graphical representation of the framework is proposed. The differences between the original framework and the revised one, and the advantages of the latter, are discussed. Some implications of the framework for the design and evaluation of information access systems and their user interfaces are also derived and, finally, an exploratory study on the issue of agreement in relevance judgments, and its consequences for the design of multimedia test collections, are presented.

Keywords: relevance, kinds of relevance, information systems design and evaluation, user interfaces, relevance judgments, test collections design.

1 Introduction

Relevance [14, 25, 26, 27] is one of the most debated and central concepts in Information Science. During the last 30 years an impressive number of studies have proposed different approaches, models, and criteria for understanding its meaning and for establishing an operational measure of it (for an exhaustive survey see [22]).

The state of the art in the field of relevance studies can raise at least two kinds of attitudes. Some people could find very disappointing the fact that, in spite of the large number of debates on relevance recently appeared, no agreed and precise definition of the concept has been reached; this could also weaken any reasonable expectation that such a definition will appear in the near future [13].

The opposite attitude toward this problem, that we want to subscribe in the present work, is to consider the point reached in relevance studies as demonstrating that this research area needs to be further investigated. Relevance is crucial for a real understanding of users information seeking behaviors. It is also important for developing appropriate methodologies for evaluating information retrieval and filtering systems (we will use *Information Access*, or IA, for referring to both information retrieval and filtering, and IAS for referring to an IA system). Evaluation studies should lead to the design of innovative systems more suited to satisfying users needs within different working contexts. Supporting the second attitude is the sensation that the lack of a precise definition of relevance is amplified by the existence of different kinds of relevance, and by some ambiguity in the terminology used. The work started in [23] and continued in this paper is an attempt to decrease this ambiguity.

In [23] a multidimensional framework for relevance has been presented. We used it as a starting point for carrying out a collaborative activity of discussion and negotiation on the concept within the IA research community participating to the MIRA workshops in 1998. On the basis of the negotiation process, we developed a new framework, that extends and refines the original one, and can be considered as more sharable because it has emerged from a collaborative activity of knowledge construction [8, 33]. We believe that the refined framework provides a useful representation of the concept of relevance since it mirrors a classification of its most important components in a way that can be judged, at present, as significant and sufficiently complete for the design and evaluation of IASs. We will also demonstrate how it can be inspiring for the development of new methodologies for multimedia test collections design.

The paper is structured as follows. We start in Section 2 by describing the negotiation activity on the issue of relevance carried out during the MIRA workshops and the shortcomings about the original framework that have emerged. In Section 3 we describe the new framework for relevance proposed in this paper. In Section 4 we provide a brief characterization of the different ways in which the meaning of information has been conceptualized in the literature, introducing then the approach to information and relevance advocated in the present work, and we emphasize some of the differences, between the original framework and the new one, that have been most debated during the negotiation and refinement process. In Section 5 we explain the usefulness of the framework for the design of innovative IASs and their interfaces. In Section 6 we discuss the issue of IASs evaluation and we describe an empirical study on relevance judgments carried out during the Dublin MIRA workshop and the Glasgow MIRA conference. Its results support our characterization of relevance and suggest some interesting implications for the design of a multimedia test collection. In Section 7 we report about some future applications and further developments of the relevance framework presented.

2 The negotiation activity on the original framework

The starting point of this research is the paper [23], in which a four dimensional framework for representing and distinguishing the various kinds of relevance has been proposed. We stimulated a negotiation activity aimed at evaluating and improving the framework presented in [23]. The negotiation activity took place in the period between two MIRA workshops held in 1998, the first one in Grenoble (March '98, http://www.dcs.gla.ac.uk/mira/workshops/grenoble) and the second one in Dublin (October '98, http://www.dcs.gla.ac.uk/mira/workshops/dublin). It consisted in the presentation and delivery of the paper [23] to some experts in information retrieval and information science belonging to the MIRA community. The 6 experts who accepted to participate to the study had to read the paper and answer to a questionnaire of 10 open ended questions. The questionnaire was designed to evaluate the completeness and comprehensibility of the framework and to raise any kind of comments, doubts, and objections on experts behalf. We analyzed and discussed the results of the survey, and then presented them for a further collective discussion during a specific session of the Dublin MIRA workshop.

The answers to the questionnaire, the comments collected during the discussion, and the subsequent analysis led us to highlight some shortcomings of the original framework:

- The first dimension, Information resources, includes entities belonging to different kinds (Document and Surrogates are specific types of information resources, Information is not).
- The concept of *metadata* is not taken into account.
- The four dimensional nature of the framework hinders, if not prevents, its graphical representation.
- The Time dimension is not conceptually structured as the other ones.
- The fourth dimension (Components) is difficult to understand and its entities constitute a partially ordered set. Moreover, it is difficult to represent graphically a partial ordering.

On the basis of these shortcomings, we elaborated the following new version of the framework.

3 The revised framework

The framework is based on a representation of relevance as a relation between 3 different dimensions: *Information Needs*, *Information Resources*, and a specific context of *Information Use*. For each dimension a series of elements have been identified.

3.1 Information Needs

Traditional studies in Information Science have agreed that what determines the event of a user searching for an information source is the experience of a 'problematic situation' or an 'Anomalous State of Knowledge' [5], inducing her to look for some kinds of information to overcome that situation.

We make a further step of abstraction, on the basis of the remark that an information need is not always perceived by a person: a teacher preparing a lesson, or a researcher writing a paper, could need some information about a new result that she does not know, yet. We represent and name this situation as a *Real Information Need* (RIN), as done in [21]. The RIN is an ideal and abstract entity which includes the complete set of information that a person should have about her problematic situation and about her need for information that could solve it appropriately.

If and when the person perceives (in a more or less correct and complete way) her RIN, she has a *Perceived Information Need* (PIN). The PIN is a representation, implicit in her mind, of both the problematic situation she is experiencing and the different means to exploit for satisfying her information need.

A person's PIN is often in a visceral and muddled form. The next step a user performs is to derive from it an *Expressed Information Need* (EIN) elaborating a request, for example, in natural language.

To interact with an IAS, a user has also to translate her EIN into a *Formalized Information Need* (FIN) using a notational representation or a language understandable by the machine, for instance using keywords and boolean operators for formulating a query.

Summarizing, the set of elements of Information Needs dimension (hereinafter InfNeeds) are represented by RIN, PIN, EIN, and FIN; we can order these elements since they form a scale of increasingly more complete representations of a user's knowledge states:

$$\text{FIN} \leq \text{EIN} \leq \text{PIN} \leq \text{RIN}.$$

The ordering is not referred to the user's ability of expressing the RIN, but to the completeness of description and potential expressiveness of it. In our view, even experts in a certain domain of knowledge have always a partial awareness of their RINs, since these are dynamic entities that evolve in time. Different kinds of information needs categorizations have been proposed by previous studies: Taylor [32] spoke of 'visceral', 'conscious', 'formalized', and 'compromised' needs; Ingwersen [18] spoke of 'verificative', 'conscious topical', and 'muddled topical' needs; and Cluzeau-Ciry [9] (see also [31]) proposed the classification of 'thematic', 'connotative', 'exploratory', and 'precise' needs for iconographic databases users. We think that the value of the representation set composing InfNeeds consists in being both complete and easy to relate to operational measures. For example if the need is expressed in natural language it is a request, if it is expressed in a formal language it is a query. The representation set is also independent of time and cause [11]. It is independent of time since the EIN may come before (or after) the PIN, as when a written request is passed by a person to another one that interprets it transforming the EIN into a PIN (and perhaps into an EIN that could be different from the original EIN). It is independent of cause, since the EIN may come from the PIN or vice versa, especially when we contemplate collaborative activities between people [11].

The elements of InfNeeds lead to consider the importance of the phenomena taking place between PIN and RIN (involving users interactions with intelligent interfaces or with human information intermediaries and experts) but also the cognitive processes taking place iteratively between the expression of an information need (EIN) and the formulation and reformulation of a specific query (FIN).

Such an ideal entity as the RIN is useful because it emphasize its potential difference from the PIN and the importance of providing users with all the facilities enhancing their understanding of what their actual information need is. For instance we can give them the opportunity of finding some kind of expert advice, both from the system used and from human resources through collaborative processes of action and communication.

The reader will be already familiar with studies and discussions about the so called 'label effect' [18] and the 'vocabulary problem' [15] which are considered as the major drawbacks observed in users interactions with textual retrieval systems. Now that IASs are characterized by multimediality, researchers attention is focusing on providing users with new ways of expressing their information needs, that can concern, for example, visual type of data. Some studies report that users prefer a more direct and modality specific way of expressing their information needs, like for instance sketch drawing or the use of images for querying a visual database [19, 16], instead of inserting textual queries, since these actions seem to be more coherent with their way of thinking about the information to look for.

Finally, the move from PIN to EIN is usually influenced by situational constraints, for instance how much time a user wants to invest in talking to an intermediary or writing down the details of her PIN; the move from EIN to FIN depends more on technology constraints, as, for instance, the number of languages and functionalities the user has at her disposal to retrieve multimedia documents and the level of difficulty implied in learning to use such facilities.

3.2 Information Resources

The second dimension constituting relevance in our framework is a classification of the types of information resources (InfRes) a user can access when experiencing a particular information need. We consider the following categories as significant in IA use:

- *Set of Documents* (DS): a certain number of documents that together can satisfy a user's information need;
- *Document* (D): a complete information object a user can access after a search;
- *Metadata* (MD): structured data about data (for instance bibliographic headers in web pages, 'terms of use' statements, information about the quality of data, etc.) but also indications or suggestions about data, provided by human resources;
- *Surrogate* (S): a representation of a document consisting in a title, a keyword, an abstract, etc.

The entities of this dimension can be ordered as follows:

$$\text{S} \leq \text{MD} \leq \text{D} \leq \text{DS}$$

on the basis of their potential capacity of providing a user with the data necessary to satisfy her RIN. For example a surrogate S can give just an indication to a user that a document can be interesting for her search goal. A metadata MD provides a richer amount of information about the utility and appropriateness of a document for the search problem but, of course, if a user can read the entire document D it is likely that she will be able to get more information to satisfy her information need. Most of the time, a RIN can only be satisfied collecting information from different sources, and that is why we indicated the higher extreme of this dimension with DS.

3.3 Information Use Context

The last dimension of the framework (abbreviated with InfUseCo) is concerned with a characterization of the context in which information is used. Context can be conceptualized as formed by the following 3 components: *topic*, *task*, and *user attributes*. This means that in a certain situation we can evaluate the relevance of an element of InfRes for an element of InfNeeds with respect to:

- Topic (To), which means the domain, the area of interest for the user, for instance 'the concept of relevance in information science'. In this case a large part of context remains implicit.
- Topic & Task (To+Ta), a more complete description of the area interesting for the user, since it considers also the activity she will perform with the retrieved documents (for instance 'the concept of relevance in information science for writing a PhD thesis in information engineering').
- A Topic, Task, & User Attributes (To+Ta+UA), that includes also other descriptions about user characteristics (such as her expertise in the domain, preferences, physical attributes, etc.) that are useful for understanding the specific context in which information is sought and retrieved.

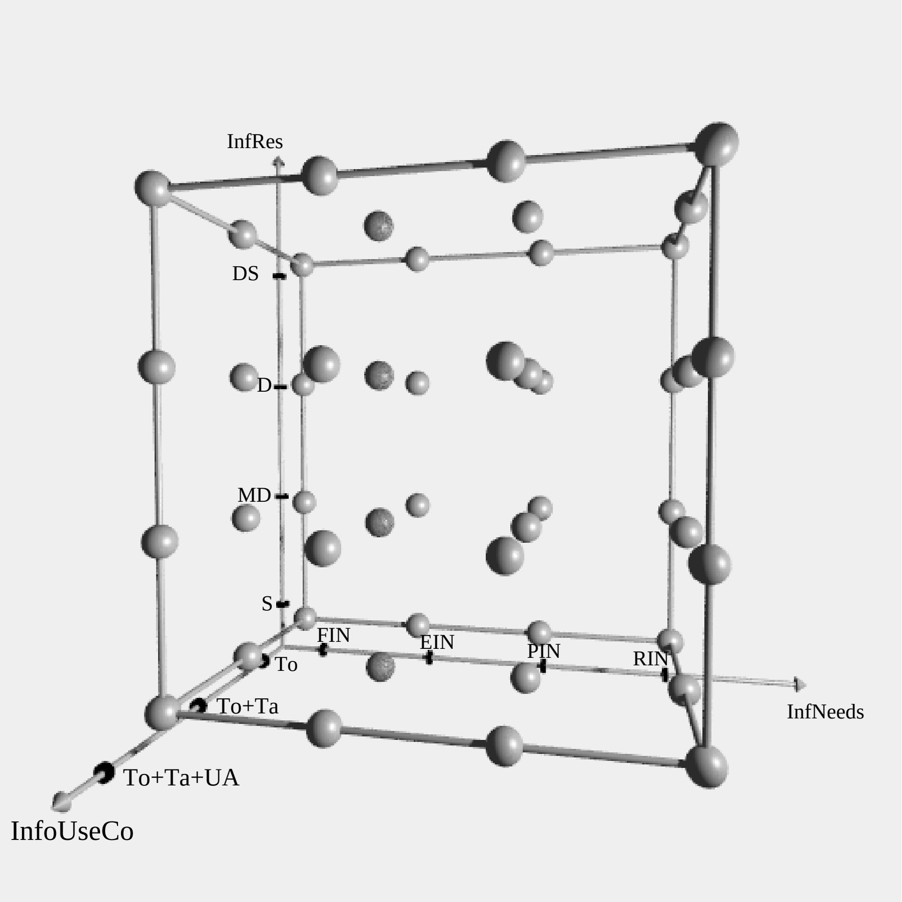


Figure 1: The three dimensional relevance space.

We represent this dimension as a set containing 3 elements, ordered as follows:

$$\text{To} \leq \text{To+Ta} \leq \text{To+Ta+UA}.$$

The different kinds of details with which we could potentially specify a context show how we could also extend our notion of relevance and our ways of measuring it going beyond pure topic consideration, for example in test collections design (see Section 6.1 below) and towards a more complete inclusion of the other activity dependent (task) and user dependent (user attributes) variables affecting information access and retrieval.

3.4 A graphical representation of relevance space

To facilitate a more intuitive understanding of the framework, we represent graphically all the possible combinations of InfNeeds, InfRes, and InfUseCo, that together constitute the three dimensional *relevance space*.

In Figure 1 we provide a cartesian representation of the framework's 3 dimensions, with axis x representing the elements of InfNeeds (FIN, EIN, PIN, RIN), axis y representing the elements of InfRes (S, MD, D, DS), and axis z representing the elements of InfUseCo (To, To+Ta, To+Ta+UA). Each sphere inside the figure represents a particular kind of relevance obtained combining the elements in the 3 dimensions (for a total of 48 relevances).

Since the framework's 3 dimensions are ordered sets of elements, we can estimate how far a particular relevance is from the one interesting for the user, namely:

rel(DS, RIN, To+Ta+UA).

4 A discussion about the revised framework

In this section we discuss the modifications and some of the potential advantages of the new framework proposed.

4.1 Fundamental conceptions about information and relevance

The concept of information does not appear as a specific entity in the revised framework (the entity Information has been removed from InfRes), because it underlies all the 3 dimension of the framework. We want also to outline some of the different ways in which the meaning of information has been intended in the literature, since this helps to clarify the underlying assumptions of the approach advocated in the present study.

Information has been defined in many ways by various researchers in this century. These definitions can be classified into two groups: "hard sciences" vs. "soft sciences" information theories. The first group (in which information is usually defined in an objective way) includes the well known Shannon's information theory [30], the Algorithmic Information Theory, independently developed by Chaitin, Kolmogorov, and Solomonoff [20], and the Semantic Information Theory introduced by Bar-Hillel, Carnap, and Popper [17] and further developed by Dretske [12], Barwise and Perry [2], and Devlin [10]. Besides these "hard sciences" information theories there are others, perhaps less known, "soft sciences" approaches (in which information is usually defined in a subjective way): Bateson defined information as a difference [3, 4], Brookes proposed that "information is a small bit of knowledge" [6], and, more recently, Clancey expanded Bateson's definition of information, proposing that information is the detection of a difference that is functionally important for an agent to adapt to a certain context [8].

Both the hard and soft sciences approaches have some peculiar strengths and weaknesses: hard sciences approaches are more mature and, obviously, formalized, but fail to capture the everyday meaning of information; on the other hand, soft sciences approaches give us an account of a more "human" information, but at a less formalized level.

Another attempt to formally describe information from a subjective viewpoint has been proposed in [21], and is still under development. The basic idea is to define information in the following way: if a datum is perceived by an agent in a certain knowledge state, and this datum leads to a change of the agent's knowledge state, the information carried by that datum is the difference between the final and the initial knowledge states. In this way the subjectivity of information can be taken into account, since different agents will have different initial knowledge states and thus will interpret the same datum in different ways.

This proposal reflects our position on the subjectivity of information for human beings: instead of talking about the objective information contained in a datum D (info(D)), we should emphasize the importance of the recipient and refer to the subjective information that a datum D gives to a recipient R (info(D,R)).

The definition of relevance that can be derived from this definition of information has to be a subjective one. Relevance is intended as a concept which specifies if, in a certain situation, the coupling between an information need and some particular type of data results appropriate for an agent to pursue its intended goal.

In this work we support the view that relevance is a useful concept if considered with reference to a subjective approach to information, not with reference to a datum or a set of data in isolation. Our description of relevance is also grounded on a dynamic view of information. So information and relevance can be defined just with reference to the specific set of entities and contingencies present in a certain moment of time. The overall assumptions on which our framework is based are coherent with the situational approach recently proposed in Information Science literature [28] and they support the elaboration of a notational model that extends previous studies by providing a more complete, precise, and formally expressible definition of relevance.

4.2 The main changes

The Time dimension, present in the original framework, has been eliminated. During the study some participants expressed their difficulty in understanding the sense of Time as an independent dimension. Many of them recognized the importance of considering the dynamic evolution of InfNeeds and InfRes but they found that a Time dimension should have been conceptualized using a number of specific and significant categories as those included in the other three dimensions. A suggestion was, for example, to compose the Time dimension with categories stressing the difference between static vs. dynamic kinds of relevance. Following this discussion we preferred to refine the framework making more explicit some of the assumptions on which it is based (information as intrinsically dynamic, see Section 4.1) and representing time as the shifting from a particular point within relevance space to a new one.

Other changes were operated on some dimensions of the original framework that were judged as rather ambiguous in their meanings.

The dimension InfRes was enlarged to include DS (what is usually searched by a user, and retrieved by an IAS) and MD. The distinction between the elements S and MD seems particularly useful for the users, even if in the literature the difference between S and MD appears sometimes rather fuzzy, with the same kind of data referred to as belonging to both of the above mentioned categories [29]. We think that IAS users should be made aware of the distinction and come to appreciate and exploit more the peculiarity of MD. An increasing number of research projects are recently concerned with the production of metadata and this efforts are intended to create new collections of organized structured data about documents incorporating an added value (if compared to the more traditional types of surrogates) that should be particularly useful to satisfying information needs. Users should take into consideration this added value of MD at the moment of relevance judging. Our framework includes into the category of MD also all the information a user can get during a search not only from documents but also from other people, like colleagues asked for advice.

Another advantage of the new framework consists in having clarified the nature of context in information search and use and the number and type of factors we can consider when analyzing this dimension (InfUseCo). The components Topic, Task, and User Attributes were combined to form a *totally ordered* set of elements; in fact in information seeking situations usually a certain topic is firstly specified and then other details about the type of task and user attributes are added, since they prove to be useful to express the context more precisely. This does not mean that Task or User Attributes elements are less important than Topic, but just that, traditionally, in relevance studies and especially in test collection design, topicality has been the only contextual factor considered, for reasons of experimental design simplification. When this happens, the entities of Task and User Attributes are not absent from the user situation, they just remain implicit. The new 3 dimensional framework has also the advantage of being easier to represent graphically (Figure 1); this benefits both the communication of the model and the exploration of relevance space that we consider important for reaching a deeper understanding of its characteristics.

5 User interfaces for information access systems

Even if the primary purpose of the framework is to represent a comprehensive conceptual model of relevance, it could also be inspiring for the design of innovative IASs user interfaces. One could implement the complete framework within an IAS interface, and test empirically if such an interface would result easy to understand and really helpful for the users. A more feasible alternative would consist in implementing some specific portions of the framework into an interface, since this might start to significantly change the way the user can express her relevance judgments to the system, improving the total performance on a task. In this section we propose some more detailed examples about how this could happen.

Figure 1 could be embedded, maybe considering just a restricted set of its entities, into a user interface, and the retrieved documents plotted on it as clickable points (inside the spheres), like in ‘starfield displays’ [1]: the user can examine the documents in an area of relevance (usually, the higher relevance in the ordering) by clicking on them, and she can drag the documents from one area of relevance to another one, signifying, for instance, that a document is relevant for the topic but not for the task. In this way, the user would be provided with an interface allowing two dual approaches to information seeking: the usual ‘navigation of the information space’ approach (see for instance [7]) and the new ‘navigation of the relevance space’ one. A similar approach was suggested in [23], but it could not be implemented directly, since no concrete solution for representing the 4 dimensional framework was available.

This interface would allow some sort of relevance feedback: by dragging a document from the ‘relevant to the topic and task’ sphere to the ‘relevant to the topic’ sphere, the user communicates to the system that some features of the document make it not suited to the task at hand, even if the topic of the document is adequate (the terms in the document can be used for positive feedback, in the usual way). This kind of relevance feedback is based on binary relevance, but it is possible to devise an enhanced kind of relevance feedback. Figure 2 illustrates this idea: a user interface like the one sketched in Figure 2a would allow the usual dichotomous relevance feedback (the user can judge the document as relevant or not relevant, by clicking on a button or dragging the document into the 'relevant' folder); the interface in Figure 2b allows a multidimensional but dichotomous relevance feedback; the interface in Figure 2c allows a weighted, but unidimensional relevance feedback; and the interface in Figure 2d allows, being a combination of the previous two, a multidimensional and weighted relevance feedback.

One might also design an interface, customized for a specific category of users, allowing them to express their RIN and to ask the system to filter InfRes like surrogates, metadata (as for instance colleagues comments derived from FAQ collections), or documents. Users relevance judgments would be strongly influenced by their knowledge about the type of data (S, MD, or D) the system provides them.

Finally, let us remark that a classical IAS presents to the user a total order of the documents retrieved. This operation becomes more difficult for an IAS working on a 'multidimensional' relevance: if a document D1 is more relevant for the topic and a document D2 is more relevant for the task, which is the better one for the user? In our opinion the correct answer is "It depends (on other user and situational characteristics)", and a system providing a total ordering of the retrieved documents would not be very useful, in such a case. This could appear as a serious limitation, since user's task is a decision task: choosing which document(s) to read. Draper [11] suggests that every user would prefer a total ordering; nevertheless, nowadays systems do not have enough information for establishing an appropriate total ordering, and so we believe it is better to leave this task to the user. If and when the IASs will have a more complete model of the user attributes and tasks, a reliable total ordering will be obtainable and useful.

6 Evaluation of information access systems

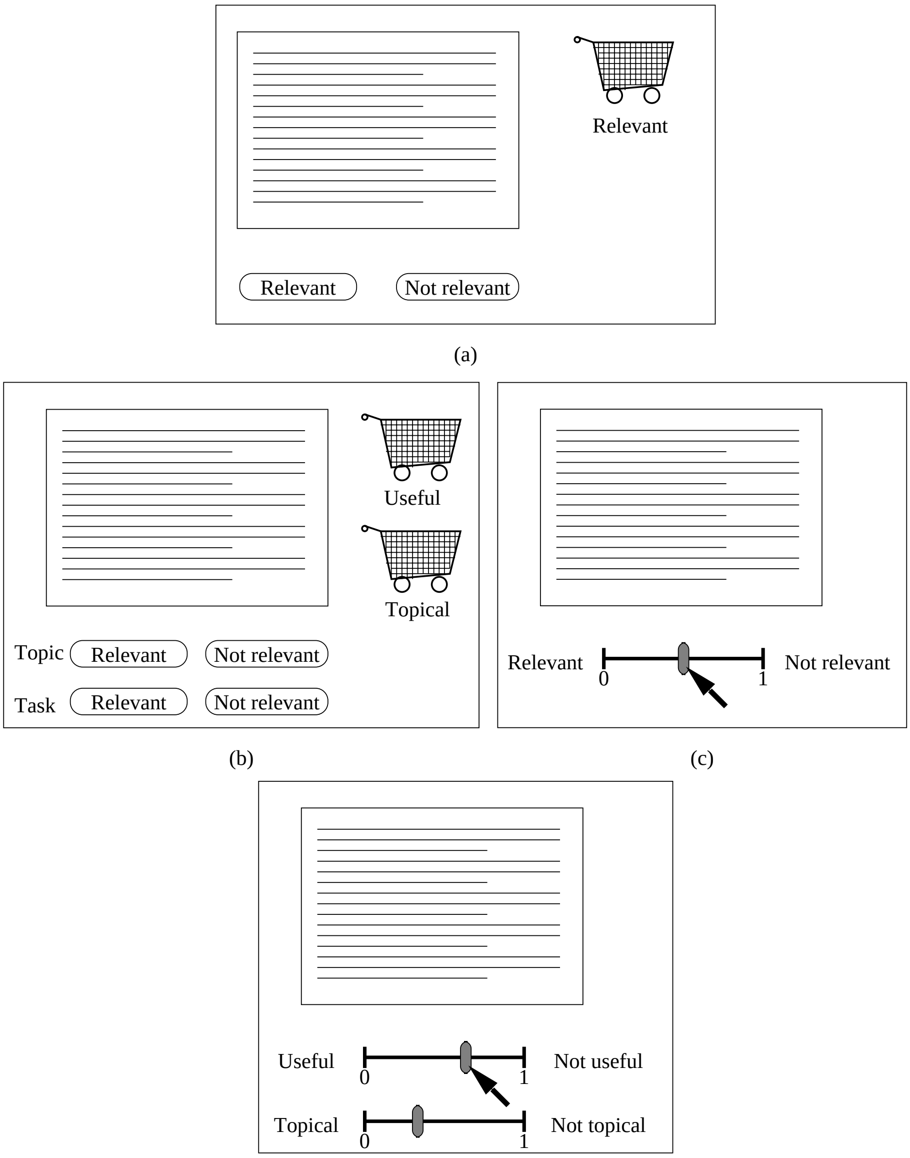
Our framework leads to conclude that relevance judgments based just on "topics" are not representative of all the other existing types of relevances, since one of the framework's three dimensions (InfUseCo) is not fully taken into account. So a more extended conception of relevance seems unquestionable: it is as going from black and white images (binary judgment) to grey levels images (scalar or weighted judgment) to color images (multidimensional judgment). But what about its practical usefulness? We believe that multidimensionality can be useful, besides for designing IASs as we have demonstrated, also for inspiring test collections construction.

6.1 Task and user attributes in test collections

It is not by chance that the queries in classical TREC test collections are called 'topics': the kind of relevance judged is topicality. The framework presented in this paper brings, in a natural way, to the proposal of building beyond-topical test collections. But the framework can be used also for handling the disagreement among judges deciding on relevance. When the disagreement on a query Q is too high, four possible strategies can be followed:

- Just calculate the mean of the judgments (in this way loosing a lot of information).
- Reject the query Q (thus not admitting into the sample the queries that lead to disagreement).
- Change Q adding a more specific topic (thus not admitting into the sample the queries with general or vague topics).
- Change Q adding task and/or user attributes.

To confirm the importance of considering beyond-topical aspects of relevance in the design of test collections, we present the results of an explorative study performed during the MIRA workshop in Dublin and the MIRA final conference in Glasgow.



(d)

Figure 2: User interfaces for relevance feedback.

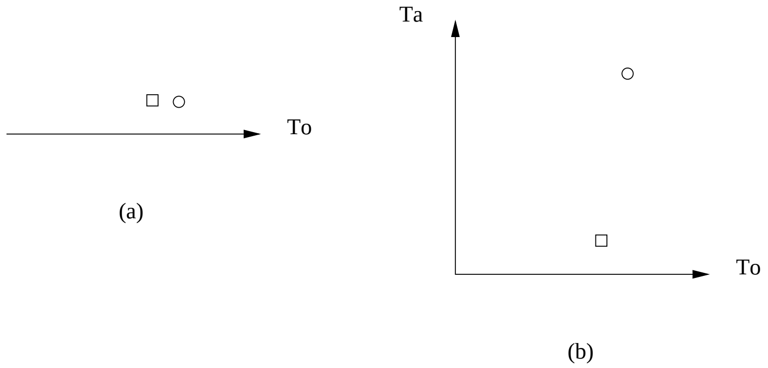


Figure 3: Adding the task might lead to a higher disagreement.

6.2 An exploratory study on the agreement in relevance judgments: Aims and motivations

After having recognized the existence of relevance space, it seems reasonable that a group of judges expressing their relevance judgments might disagree just because they refer to different kinds of relevance. Even if the judges are asked, or expected, to judge the same kind of relevance, a problem might arise. For instance, if the task is not explicitly stated in the query, each judge might refer to a different, implicitly constructed, task. If the judges are "similar", the implicitly constructed tasks may be similar, but if the judges are different for culture, education, and so on, the tasks may vary a lot, and judgments too. The same considerations hold, of course, for user attributes. Therefore, adding task or user attributes to the query may reduce the ambiguity, and decrease the disagreement between judges.

If the task is not expressed in the query, the judges might just disregard it, and judge the relevance for the topic component only. In this case, the disagreement can significantly increase when a task is explicitly added to the query, if this added task is quite ambiguous and if it is interpreted in different ways by the judges. This is not surprising at all: adding the task is like adding another dimension, and, obviously, two points (or documents) that appear very close in a one-dimension view (see Figure 3a) can actually be quite far when another dimension is taken into account (see Figure 3b).

Finally, note that the study of human relevance judgments is becoming increasingly important as new media are being used: human relevance judgments are more easily obtainable for images than for text, since it is far quicker to judge the relevance of an image than a text. To investigate more on these issues, we first designed and performed the following experiment.

6.3 Experimental setting, data analysis, and results of the study

We asked to the Dublin workshop participants to imagine of being a young woman interacting with an online catalog presenting fashion photos, and to express some relevance judgments about them.

The participants were asked to rank 5 photographs of dresses in descending order of relevance for 8 queries (Figure 4). The 8 queries had different levels of detail: the first one contained only a topical context, while the other ones added a task, some user attributes, or both.

The participants involved in the experiment were 26; 7 protocols were discarded for data missing problems, obtaining 19 valid protocols each containing the expression of 8 total orderings, one for each query.

We were interested in measuring the average disagreement both between different judges for the same query and within the same judge for different queries. For measuring the disagreement between two orderings we chose to use the distance defined as "the minimum number of switches of adjacent elements", so that the distance between (12345) and (12354) is 1; between (12345) and (12543) is 3, because (12345) → (12354) → (12534) → (12543); and between (12345) and (54321) is 10 (as the reader can verify), and this is the maximum distance between two orderings, ie., the maximum disagreement between two judges. These values are then normalized in the $[0..1]$ range, dividing each of them by the maximum distance obtainable (10 in the case of orderings for 5 documents); we also use the *group*

The five dresses:

1. A simple grey dress.
2. A simple violet dress.
3. An elegant long dress.
4. An elegant short dress.
5. A Carnival costume.

The eight queries:

1. Find a dress to look smart. [Topic]
2. Find a dress to look smart for participating to a wedding [Topic + Task]
3. Find a dress to look smart for going to my office. [Topic + Task]
4. Find a dress to look smart for going to do some shopping. [Topic + Task]
5. Find a dress to look smart for going to the Venice Carnival. [Topic + Task]
6. Find a dress to look smart even if my legs are not perfect. [Topic + User Attributes]
7. Find a dress to look smart that fits with my red hairs. [Topic + User Attributes]
8. Given that my legs are perfect, find a dress to look smart for going to a wedding (‘my legs are perfect’, she thinks, but this is not true). [Topic + Task + User Attributes]

Figure 4: Documents (dresses photos) and queries for the experiment.

disagreement, that is the average of the disagreement between two judges in a group, again normalized since it tends to decrease with the number of judges involved (see [24] for more details).

The results are summarized in Table 1 (the average distance of two orderings for the first query is 0.24; for the second it is 0.33; and so on) and Table 2 (the average distance of two random orderings by the first judge is 0.43; for the second judge it is 0.40; etc.).

The group disagreement in Table 2 (0.22) is the double of the group disagreement in Table 1 (0.11). This means that adding task and user attributes to the query causes less disagreement than the subjectivity of the judges.

The average distance for query number 1 (Table 1) is 0.24, quite close to the mean. At a first glance, this seems a negative result, since it contrasts with the expectation that, being a topic-only query, the disagreement on it was expected to be higher. Nevertheless, this can be seen under a different light by analyzing more carefully the data in Table 1, that show a rather high difference among the various queries (confirmed by the high standard deviation): queries 3, 4, and 5 have a low disagreement (about 0.1), while queries 2 and 7 a high one (more than 0.3). This phenomenon can be understood looking at the differences among the queries:

- The task added in query number 2 is quite ambiguous, since there are many kinds of wedding ceremonies all other the world: we had participants from various European countries (Denmark, Finland, France, Germany, Great Britain, Ireland, Italy, Turkey) with different cultural education, religion, etc. So, the disagreement on this query is not very surprising.

- The task added in query number 7 refers to the fact, probably ignored by many participants, that a violet dress is not adequate for a woman with red hair.

Therefore, queries 2 and 7 seem to be a manifestation of the phenomenon described in Section 6.2, in Figure 3.

On the basis of these results, we can derive that adding task and user attributes to the query can be useful in many cases, since it can lead to a lower disagreement among judges. However, it can also be dangerous in other cases, since it can generate further subjectivity. So this situation should be carefully considered, before taking any decision.

During the final MIRA Conference in Glasgow we carried out another experimental study on the agreement in relevance judgments to extend the Dublin experiment. The aims were to explore the influence level of UA on relevance judgments, and to compare the obtained results with the ones emerged from the Dublin study.

Query	**Distance**
1	0.24
2	0.33
3	0.11
4	0.11
5	0.10
6	0.22
7	0.36
8	0.27
Average = 0.22	
Standard deviation = 0.10	
Group disagreement = 0.11	

Table 1: Average distance between two judges for the same query.

Subject	**Distance**
1	0.43
2	0.40
3	0.36
4	0.43
5	0.38
6	0.44
7	0.44
8	0.46
9	0.46
10	0.46
11	0.38
12	0.55
13	0.44
14	0.46
15	0.41
16	0.42
17	0.40
18	0.41
19	0.36
Average = 0.43	
Standard deviation = 0.04	
Group disagreement = 0.22	

Table 2: Average distance between two queries for the same subject.

The ten persons:

1. Naomi Campbell (top model)
2. Kim Basinger (actress)
3. Sean Connery (actor)
4. Brad Pitt (actor)
5. William Shakespeare (writer)
6. Rita Levi Montalcini (scientist)
7. Woody Allen (comic actor)
8. Whoopi Goldberg (comic actress)
9. Hilary Clinton (politician)
10. Kofi Annan (politician)

The eight queries:

1. Who would you prefer?
2. Who would you prefer to spend one hour with?
3. Who would prefer to share your house with?
4. Who would you prefer to go out for dinner with?
5. Who would you prefer to have as a colleague?
6. Who would you prefer to go to the theater with?
7. Who is the most intelligent?
8. Who is the most beautiful/handsome?

Figure 5: Documents (photos) and queries for the second experiment.

We asked the participants to answer 8 questions expressing an ordering of 10 photos of famous persons (see Figure 5). This time, participants had to express their judgments in first person, to avoid possible misinterpretations of the task assigned (as noticed in a few cases in Dublin), and to collect judgments based on their real personal attributes. Some preliminary results, based on 26 valid protocols, are presented in Tables 3 and 4 and resume the outcome of the same data analysis applied to the Dublin experiment.

The introduction of UA has produced a much higher level of disagreement between judges for a same query than the one obtained in Dublin (group disagreement 0.21 in Glasgow vs. 0.11 in Dublin); moreover the group disagreement value is very close to the group disagreement values expressing judges subjectivity obtained both in Dublin (0.22, Table 2) and in Glasgow (0.17, Table 4).

An interpretation of these results can be that there is a strong influence of individual differences when the task situation analyzed includes the consideration of UA and this is exemplified by the typical case in which users of an IA system express their personal relevance judgments on the retrieved documents after a search.

There is no significant difference between the inter-judge distances measured from query 1 to query 6 (where the queries involved just UA or both UA and the indication of a specific kind of task to be performed). However, we noticed a dramatic drop in participants disagreement on queries 7 and 8 (about half the value of the other distances). This result can be understood analyzing the nature of queries 7 and 8: basically they ask judges to refer to the abstract conceptions of ‘beauty’ and ‘intelligence’ that can be considered as culturally shared constructs within a community of people, homologating the expression of judgments about them. We have to take into account also that the sample was mainly composed by individuals belonging to the same professional world of academic research and that abstract, value laden conceptions are usually shared by larger numbers of individuals than specific ritual practices [8, 33], such as the wedding ceremonies we considered for example in query 7 of the Dublin experiment.

We are currently carrying out further analyses on the data of this second experiment, but the main lessons learned

Query	Distance
1	0.48
2	0.44
3	0.48
4	0.46
5	0.41
6	0.43
7	0.24
8	0.23
Average = 0.40	
Standard deviation = 0.10	
Group disagreement = 0.21	

Table 3: Average distance between two judges for the same query.

Subject	Distance
1	0.46
2	0.16
3	0.30
4	0.39
5	0.34
6	0.36
7	0.33
8	0.19
9	0.45
10	0.40
11	0.38
12	0.30
13	0.31
14	0.40
15	0.19
16	0.39
17	0.25
18	0.36
19	0.48
20	0.40
21	0.43
22	0.39
23	0.32
24	0.21
25	0.25
26	0.21
Average = 0.33	
Standard deviation = 0.09	
Group disagreement = 0.17	

Table 4: Average distance between two queries for the same subject.

so far seem to be: (i) the importance of acknowledging the pervading effect of UA in everyday expression of relevance judgments, and (ii) the potential usefulness of exploiting the 'culture' variable when the problem is to reduce as much as possible the disagreement between judges.

However we also believe that this kind of problems have no clear solution because they represent some still open research questions that need to be further investigated during multimedia test collections design activities. They have started to be discussed during the last MIRA meetings, where some ways of facing them have been proposed: for example it has been suggested to observe and experiment more on the process that leads different judges to reach an agreement about the relevant documents for a query as a result of a group discussion about it [11].

7 Conclusions and future work

In this work we have contributed to explain the multidimensionality of relevance by presenting a conceptual and graphical representation of its space. We have also considered the implications of using this framework as an interesting tool for implementing a new kind of IA user interface, enabling for the first time a user to navigate relevance space. The framework has also been proposed as an inspiring conceptual model from which deriving some research questions and guidelines for the design of multimedia test collections.

We are presently working on transforming the graphical representation of the framework (Figure 1) in a relevance virtual environment using a virtual reality software (we are experimenting with VRML and Java3D) easily accessible through the Web. The virtual environment would enable a user to navigate the relevance space and to interact with it, like, for instance, pointing at and selecting each specific kind of relevance for retrieving some useful information about it. Some examples could be: a formal description of its elements, a description of a prototypical situation representing that type of relevance, some bibliographical references on studies that have enquired about it, etc. We believe that the possibility of exploring and interacting with a representation of the framework could carry the benefits of making easier to learn and discuss the framework's structure and contents within a community of experts, but could also support its further development, application, and experimental evaluation in IA situations.

Acknowledgments

We would like to thank the six experts who participated to the relevance framework study (Nathalie Denos, Steve Draper, Massimo Melucci, Jane Reid, Eero Sormunen, and Alan Smeaton's IR group in Dublin) and all the participants to the discussions and experiments at the MIRA meetings in Dublin and Glasgow who gave us some useful insights. We also want to thank Andrea Fusiello for some hints on the *Rayshade* software. The work of S. Mizzaro was supported in part by a grant from the Italian National Research Council (grant nr. 210.15.11).

References

[1] C. Ahlberg and B. Shneiderman. Visual information seeking: Tight coupling of dynamic query filters with starfield displays. In *ACM CHI '94 Conference Proceedings*, pages 313–317, Boston, MA, 1994.

[2] J. Barwise and J. Perry. *Situations and attitudes*. MIT Press, Cambridge, MA, 1983.

[3] G. Bateson. *Steps to an ecology of mind*. Ballantine Books, New York, 1972.

[4] G. Bateson. *Mind and Nature — A Necessary Unity*. Dutton, E. P., 1979.

[5] N. J. Belkin, R. N. Oddy, and H. M. Brooks. ASK for information retrieval: Part I. Background and theory. *Journal of Documentation*, 38(2):61–71, 1982.

[6] B. C. Brookes. The foundations of information science. Part I. Philosophical aspects. *Journal of Information Science*, 2:125–133, 1980.

[7] I. Campbell and C.J. van Rijsbergen. The ostensive model of developing information needs. In P. Ingwersen and N. O. Pors, editors, *Information Science: Integration in Perspective — Proceedings of CoLIS2*, pages 251–268, Copenhagen, Denmark, October 1996. The Royal School of Librarianship.

[8] W. J. Clancey. *Situated Cognition—On Human Knowledge and Computer Representations*. Cambridge University Press, Cambridge, UK, 1997.

[9] M. Cluzeau-Ciry. Typologie des utilisateurs et des utilisations d'une banque d'images. *Le Documentaliste*, 25(3):115–120, 1988.

[10] K. Devlin. *Logic and Information*. Cambridge University Press, Cambridge, England, 1991.

[11] S. Draper. Mizzaro's framework for relevance [WWW document]. URL http://www.psy.gla.ac.uk/˜steve/stefano.html, 16 August 1998, August 14. (visited 29 December 1998).

[12] F. Dretske. *Knowledge and the Flow of Information*. Bradford Books, MIT Press, 1981.

[13] T. J. Froehlich. Relevance reconsidered—Towards an agenda for the 21st century: Introduction to special topic issue on relevance research. *Journal of the American Society for Information Science*, 45(3):124–133, April 1994.

[14] T. J. Froehlich (editor). Special topic issue: Relevance research. *Journal of the American Society for Information Science*, 45(3), April 1994.

[15] G. W. Furnas, T. K. Landauer, L. M. Gomez, and S. T. Dumais. The vocabulary problem in human-system communications. *Communications of the Association for Computing Machinery*, 30(11):964–971, 1987.

[16] S. Gabrielli. Social and cognitive factors in the design and evaluation of multimedia systems. In *20th Annual BCS-IRSG Colloquium on IR Research*, Springer's Electronic Workshops in Computing series publication, Grenoble, 1998. Springer. http://www.ewic.org.uk/.

[17] J. Hintikka. On semantic information. In J. Hintikka and P. Suppers, editors, *Information and Inference*, pages 3–27. D. Reidel publishing company, Dordrect-Holland, 1970.

[18] P. Ingwersen. *Information Retrieval Interaction*. Taylor Graham, London, 1992.

[19] J. J. Joemon, J. Furner, and D. J. Harper. Spatial querying for image retrieval: A user-oriented evaluation. In *Proceedings of the 21st ACM-SIGIR*, pages 232–240, Melbourne, Australia, 24-28 August 1998.

[20] M. Li and P. Vitanyi. *An Introduction to Kolmogorov Complexity and Its Applications*. Springer Verlag, New York, 1997.

[21] S. Mizzaro. A cognitive analisys of information retrieval. In P. Ingwersen and N. O. Pors, editors, *Information Science: Integration in Perspective — Proceedings of CoLIS2*, pages 233–250, Copenhagen, Denmark, October 1996. The Royal School of Librarianship. Paper awarded with the "CoLIS2 Young Scientist Award".

[22] S. Mizzaro. Relevance: The whole history. *Journal of the American Society for Information Science*, 48(9):810–832, September 1997. John Wiley & Sons Inc., New York, NY. Republished in "Historical Studies in Information Science", T. Bellardo Hahn e M. Buckland editors, 1998, ISBN:1-57387-062-5.

[23] S. Mizzaro. How many relevances in information retrieval? *Interacting With Computers, Elsevier, The Netherlands*, 10(3):305–322, June 1998. ISSN: 0953-5438. Paper awarded with the Informer (British Computer Society IR Group newsletter) 'Best Student Paper in IR'.

[24] S. Mizzaro. Measuring the agreement among relevance judges. In this volume, 1999. A preliminary version is available as a Research report of the Department of Mathematics and Computer Science, University of Udine, Via delle Scienze, 206 — Loc. Rizzi — Udine, Italy, report nr. UDMI/05/99.

[25] T. Saracevic. Relevance: A review of and a framework for the thinking on the notion in information science. *Journal of the American Society for Information Science*, 26(6):321–343, 1975.

[26] T. Saracevic. Relevance reconsidered '96. In P. Ingwersen and N. O. Pors, editors, *Information Science: Integration in Perspective — Proceedings of CoLIS2*, pages 201–218, Copenhagen, Denmark, October 1996. The Royal School of Librarianship.

[27] L. Schamber. Relevance and information behavior. In *Annual Review of Information Science and Technology*, volume 29, pages 3–48. 1994.

[28] L. Schamber, M. B. Eisenberg, and M. S. Nilan. A re-examination of relevance: Toward a dynamic, situational definition. *Information Processing & Management*, 26(6):755–776, 1990.

[29] P. Schauble and A. F. Smeaton (editors). A research agenda for digital libraries,. Summary Report of the Series of Joint NSF-EU Working Groups on Future Directions for Digital Libraries Research, 12 October 1998.

[30] C. E. Shannon. A mathematical theory of communication. *Bell System Tech. Journal*, (27):379–423, 623–656, 1948. http://cm.bell-labs.com/cm/ms/what/shannonday/paper.html.

[31] B. Simonnot and M. Smail. Model for interactive retrieval of videos and still images. In Nwosu, Berra, and Thuraisingham, editors, *Multimedia Database Systems - Design and Implementation Strategies*, chapter 10, pages 278–297. Kluwer Academic Publishers, 1996.

[32] R. S. Taylor. Question-negotiation and information seeking in libraries. *College and Research Libraries*, 29:178–194, 1968.

[33] E. Wenger. *Communities of practice: Learning, meanings, and identity*. Cambridge University Press, New York, 1998.