How many relevances in information retrieval?

Stefano Mizzaro

Department of Mathematics and Computer Science
University of Udine
Via delle Scienze, 206 — Loc. Rizzi — 33100 Udine — Italy
Ph: +39 (432) 55.8456 — Fax: +39 (432) 55.8499
E-mail: mizzaro@dimi.uniud.it
WWW: http://www.dimi.uniud.it/~mizzaro

Abstract

The aim of an information retrieval system is to find relevant documents, thus relevance is a (if not 'the') central concept of information retrieval. Notwithstanding its importance, and the huge amounts of research on this topic in the past, relevance is not yet a well understood concept, also because of an inconsistently used terminology. In this paper, I try to clarify this issue, classifying the various kinds of relevance. I show that: (*i*) there are many kinds of relevance, not just one; (*ii*) these kinds can be classified in a formally defined four dimensional space; and (*iii*) such classification helps us to understand the nature of relevance and relevance judgement. Finally, the consequences of this classification on the design and evaluation of information retrieval systems are analysed.
Keywords: Information retrieval, relevance, kinds of relevance, relevance judgement, system design, system evaluation.

1 Introduction

Relevance is a crucial concept of *Information Retrieval* (IR) (Ingwersen, 1992; Salton, 1989; van Rijsbergen, 1979), as the aim of an IR system is to find *relevant* documents. Many researchers studied this issue; the main events (for an extensive survey see Mizzaro, 1997b) probably are:

Vickery (1959a; 1959b) presents at the 1958 ICSI debate a distinction between 'relevance to a subject', that refers to what the IR system judges 'relevant', and 'user relevance', that refers to what the user needs.

Rees and Schultz (1967) experimentally study the effect of different scaling techniques on the reliability of judgements. They note that relevance judgements are inconsistent and affected by about 40 variables.

Cuadra and Katter (1967a; 1967b; 1967c) experimentally find 38 variables (for instance: style, specificity, and level of difficulty of documents) that affect the relevance judgement, thus questioning the reliability of human relevance judgement.

Cooper (1971) defines relevance on the basis of notions borrowed from mathematical logic, namely entailment and minimality. A sentence s is defined to be relevant to another sentence r (or to its logical negation $\neg r$) if s belongs to a minimal set of premises M entailing r. In symbols: *relevant(s,r)* **iff** $\exists\ M\ (s \in M \wedge M \models r \wedge M - s \not\models r)$. Then, a document D is seen as a set of sentences $D = \{s_1, s_2, \ldots, s_n\}$, and its relevance to a request r is defined as: *relevant(D,r)* **iff** $\exists\ i$ *(relevant*(s_i, r)*)*.

Wilson (1973) tries to improve Cooper's definition, and uses the term *situational relevance*. He introduces the 'situation', the 'stock of information', and the goals of the user, and claims that probability and inductive logic, in addition to the deductive one used by Cooper, have to be used in defining relevance.

Saracevic (1975) reviews the papers on relevance published before 1975, and proposes a framework for classifying the various notions of relevance proposed until then.

Schamber, Eisenberg, and Nilan (1990) join the user-oriented (as opposed to the system-oriented) view of relevance. They maintain that relevance is a multidimensional, cognitive, and dynamic concept, and that it is both systematic and measurable.

Froehlich (1994) introduces the special topic issue of the Journal of the American Society for Information Science on the topic of relevance (JAS, 1994), listing six common themes of the papers in that issue: (1) inability to define relevance; (2) inadequacy of topicality; (3) variety of user criteria affecting relevance judgement; (4) the dynamic nature of information seeking behavior; (5) the need for appropriate methodologies for studying the information seeking behavior; and (6) the need for more complete cognitive models for IR system design and evaluation.

Schamber (1994) reviews, in the first ARIST chapter devoted entirely to relevance, the literature on relevance (concentrating on the period 1983–1994) and proposes three fundamental themes and related questions: (1) Behavior (What factors contribute to relevance judgments? What processes does relevance assessment entail?); (2) Measurement (What is the role of relevance in IR system evaluation? How should relevance judgment be measured?); and (3) Terminology (What should relevance, or various kinds of relevance, be called?).

Notwithstanding the importance of relevance, and the efforts made since the 60s for understanding its nature, still today it is not a well understood concept. In my opinion, and according to the above mentioned third fundamental theme in (Schamber, 1994), a great deal of such problems are caused by the existence of *many* relevances, not just *one*, and by an inconsistently used terminology: the terms 'relevance', 'topicality', 'utility', 'usefulness', 'system relevance', 'user relevance', and others are given different meanings by different authors; sometimes they are used as synonyms (two or more terms for the same concept) and sometimes in an ambiguous manner (the same term for two or more concepts).

In this paper (that revises, refines, extends, and formalises previous work, see Mizzaro, 1995; 1996b; 1996c), I try to put in order this "relevance pool"

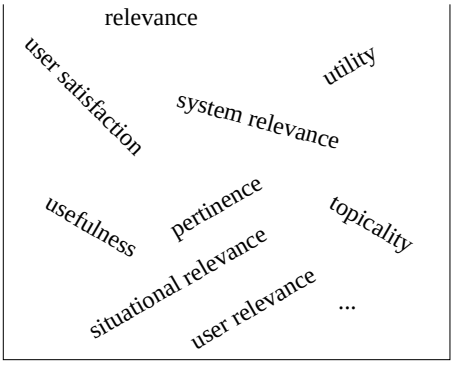

Figure 1: The "Relevance pool".

(Figure 1): Section 2 describes a framework for classifying the many kinds of different relevances in a formalised four-dimensional space; Section 3 discusses the issue of relevance *judgement*; Section 4 shows how to use the framework for analysing the literature on relevance and presents the consequences of the classification on the implementation and evaluation of IR systems; and Section 5 concludes the paper. In this paper there are some very simple mathematical formulas that, I think, help to have a clear and schematic style. Anyway, the reader can simply skip them without problems.

2 The four dimensions of relevance

In order to characterise the various kinds of relevance, in the following four subsections I define four ordered sets, each being one dimension for the relevance classification; then, Section 2.5 presents the various kinds of relevance.

2.1 First dimension: information resources

It is commonly accepted (Lancaster, 1979) that relevance is a relation between two entities of two groups. In the first group, we can have one of the following three entities:

- *Document*, the physical entity that the user of an IR system will obtain after his seeking of information;
- *Surrogate*, a representation of a document, consisting of one or more of the following: title, list of keywords, author(s) name(s), bibliographic data, abstract, and so on;
- *Information*, the (not physical) entity that the user receives/creates when reading a document.

Therefore, the first set is the *set of information resources*:

$$InfRes = \{Surrogate, Document, Information\},$$

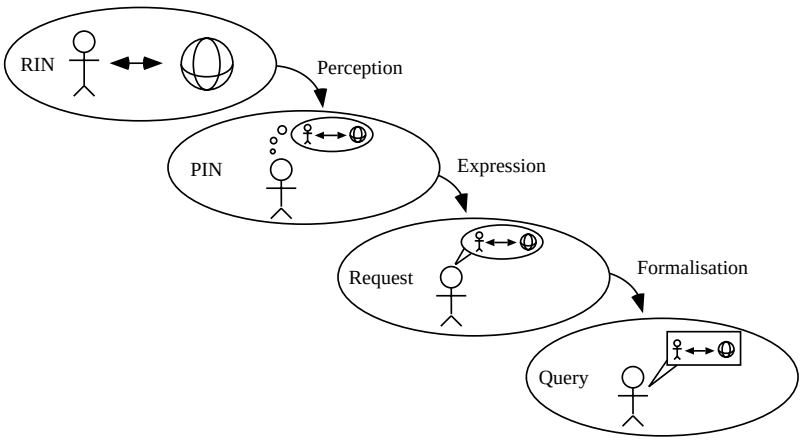

Figure 2: Real information need, perceived information need, request, and query.

that can be *ordered* as follows:

$$Surrogate < Document < Information.$$

This order might be justified on the basis of the information carried by an entity of *InfRes*, but its utility will be clear at the end of Section 2.2.

2.2 Second dimension: representation of the user's problem

The entities of the second group appear if one analyses the interaction between a user and an IR system. The user is in a "problematic situation" (Belkin et al., 1982a; 1982b), that needs information for being solved: he has a *need* of *information*, usually named, in the IR field, *information need*. Actually, for reasons that will be evident immediately, I prefer to call it *Real Information Need* (RIN).

The user *perceives* the RIN and builds the *Perceived Information Need* (PIN), a representation (implicit in the mind of the user) of the problematic situation. It is different from the RIN, as the PIN is a mental representation of the RIN; moreover, the user might not perceive in the correct way his RIN.[1]

Then the user *expresses* the PIN in a *request*, a representation of the PIN in a 'human' language, usually in natural language, and finally he *formalises* (perhaps with the help of an intermediary) the request in a *query*, a representation of the request in a 'system' language, *eg.* boolean. These four entities (RIN, PIN, request, and query) and three operations (perception, expression, formalisation) are graphically represented in Figure 2.

So the second set, the *set of representations of user's problem* is

$$Repr = \{RIN, PIN, Request, Query\},$$

[1]See also (Mizzaro, 1996a; 1997a) for a formal definition of RIN and PIN.

and it can be ordered as follows:

$$Query < Request < PIN < RIN.$$

Taylor (1968) proposed a similar framework. He used the terms:

- *Visceral need*: the actual but unexpressed need of information, corresponding to the RIN;
- *Conscious need*: the conscious description of the need, corresponding to the PIN;
- *Formalised need*: the expression of user's need, as it would be expressed without the presence of the IR system;
- *Compromised need*: the expression of user's need as presented to the IR system, corresponding to the request.

The elements of the *Repr* set proposed here are quite similar to the four levels proposed by Taylor, and the choice of one alternative is more a matter of opinion than an objective issue. Anyway, as it will be clear later, the actual choice of the elements of *Repr* is not important: the crucial point is that there are various and different levels of representation of the user's problem.

Note that the three operations are not so simple as it might seem at first glance, since some well-known problems may appear:

- The perception operation (from RIN to PIN) is difficult as the user can be in a problematic situation: he has to search something that he does not know. This issue has been named in various ways by different researchers:
  - Mackay (1960) spoke of "incompleteness of the picture of the world" and of "inadequacy in what we may call his [the agent's] 'state of readiness' to interact purposefully with the world";
  - Belkin et al. (1982a; 1982b) introduced the acronym ASK (Anomalous State of Knowledge), for emphasising that the user might not know what he wants to know;
  - Ingwersen (1992) coined the ASK-like acronyms ISK (Incomplete State of Knowledge) and USK (Uncertain State of Knowledge), unifying ASK, ISK and USK in a common concept.
- The expression operation is hindered by:
  - the *label effect* (Ingwersen, 1992), an experimentally verified behaviour of the user that expresses his need in terms of "labels", or keywords, and not as a complete statement;
  - the *vocabulary problem* (Furnas et al., 1987), the mismatch between the terms used in the documents and the terms used in the request, derived from the existence of ambiguous terms and synonyms in natural language.
- The formalisation operation is not simple since the language used (the 'system language', not the 'human one') can be not easily understandable by the user.

Because of these problems, usually there is only a partial translation of the RIN into the PIN, then into the request and finally into the query.

As an example, let us imagine that a university professor gives to a student the homework of writing a ten pages paper about, say, "practical applications of artificial intelligence to information filtering". And let us suppose that the student has a very confused idea of what information filtering is, so that he misunderstands the topic and looks for "practical applications of artificial intelligence to information *retrieval*". He does not know anything about this topic, so he decides to query a database for finding some documents. He speaks with the intermediary and expresses his need simply saying "I need documents about applications of artificial intelligence to information retrieval". The intermediary finally submits the boolean query "artificial intelligence AND information retrieval".

In this case the query is "artificial intelligence AND information retrieval", the request is "documents about *applications* of artificial intelligence to information retrieval", the PIN is "documents about *practical* applications of artificial intelligence to information retrieval", and the RIN is "first of all, documents explaining what information filtering is, and then documents about practical applications of artificial intelligence to information filtering".[2] Query, request, PIN, and RIN are obviously quite different: documents that seem useful for the query can be absolutely not useful for the RIN, and vice-versa.

On this basis, a relevance can be seen as a relation between two entities, one from each group: the relevance of a surrogate to a query, or the relevance of the information received by the user to the RIN, and so on. Therefore, a relevance seems a point in a two-dimensional space, as illustrated in Figure 3: on the horizontal axis are the elements of *Repr*, on the vertical axis are the elements of *InfRes*; each relevance is represented by a white circle, and the arrows represent a partial order among the relevances. This order is induced by the orders of the sets *Repr* and *InfRes* and denotes how much a relevance is near to the relevance of the information received to the RIN, the one to which the user is interested, and how is difficult to measure it: the number of steps (arrows) needed to reach the topmost and rightmost circle is an indication of the distance of each kind of relevance from the relevance of the information received to the RIN.

At this point, we could classify some well known relevances, for instance:

- Vickery's (1959a; 1959b) 'relevance to a subject' and 'user relevance' could be the lower left and upper right circle, respectively;
- The relevance used in the classical IR evaluation experiments, namely Cranfield (Cleverdon et al., 1966) and TREC (Harman, 1993), is again a "lower" one (the lower left circle, or the slightly "higher" relevance of a surrogate to the request).

But the above described relevances are not all the possible relevances, since two more dimensions have to be taken into account.

2.3 Third dimension: time

The third dimension is the time: a surrogate (a document, some information) may be not relevant to a query (request, PIN, RIN) at a certain point of time,

[2] Of course, these are just *representations* of PIN and RIN.

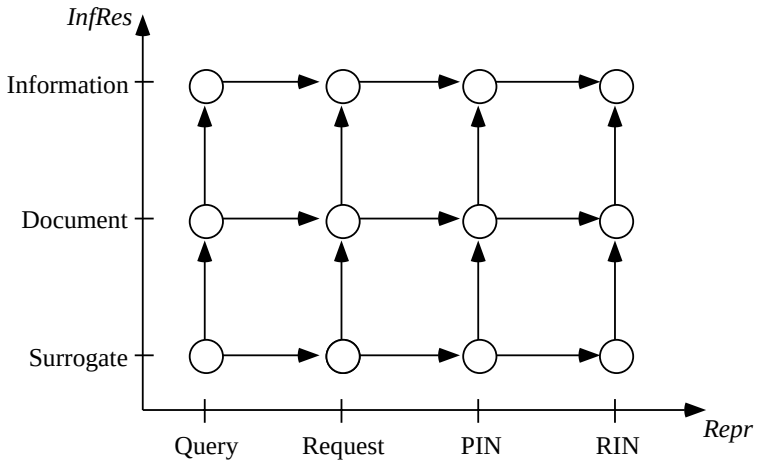

Figure 3: Relevance as a point in a two-dimensional space.

and be relevant later, or vice versa. This happens, for instance, if the user learns something that allows him to understand a document, or if the user RIN changes, and so on. Thus the scenario represented in Figure 2 has to be improved in order to take into account the highly dynamic interaction between the user and the IR system. In Figure 4 the transformations of RIN, PIN, request, and query are illustrated. Four levels, represented by the four ellipses, can be individuated, and refer to these four elements. At time $t(rin_0)$ the user has a RIN rin_0. The user perceives it, obtaining the initial PIN (pin_0, at time $t(pin_0)$), he expresses it, obtaining the initial request (r_0, at time $t(r_0)$), and he formalises it, obtaining the initial query (q_0, at time $t(q_0)$). Then a revision takes place: the initial query may be modified (obtaining q_1, at time $t(q_1)$), the same may happen for the request and the PIN, until the final information need (pin_p, at time $t(pin_p)$), request (r_m, at time $t(r_m)$), and query (q_n, at time $t(q_n)$) are obtained.

The set for this third dimension is the *set of the time points* from the arising of the user's RIN until its satisfaction:

$$\begin{aligned} Time \quad = \quad & \{t(rin_0), t(pin_0), t(r_0), t(q_0), t(q_1), t(r_1), t(q_2), \ldots, \\ & t(pin_p), \ldots, t(r_m), \ldots, t(q_n), t(f)\}, \end{aligned}$$

with the order, induced by the normal temporal order, obtained reading the elements from left to right.[3]

2.4 Fourth dimension: components

The fourth and last dimension is a bit more complex, since the previous three sets were totally ordered, while this fourth set is only partially ordered.

[3] Again, the choice of the time points used is not important, as the focus is on the time dependency of RIN, PIN, request, and query, and the ordering usefulness will be clear later, in Section 2.5.

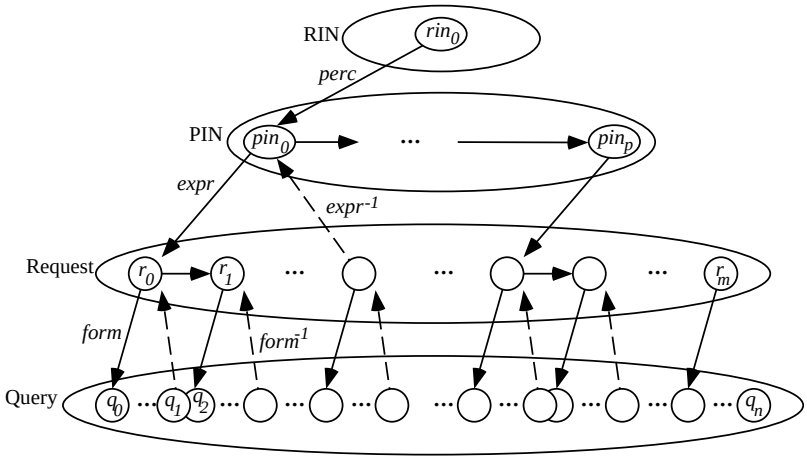

Figure 4: The dynamic interaction user-IR system.

The above mentioned entities of the first two dimensions can be decomposed into the following three *components* (fourth dimension) (Brajnik et al., 1996):

- *Topic*, that refers to the subject area interesting for the user. For example, 'the concept of relevance in information science';
- *Task*, that refers to the activity that the user will execute with the retrieved documents. For example: 'to write a survey paper on ...', or 'to prepare a lesson on ...';
- *Context*, that includes everything not pertaining to topic and task, but however affecting the way the search takes place and the evaluation of results. For example, documents already known, or not understandable, by the user (and thus not worth being retrieved), time and/or money available for the search, and so on. This component could be further decomposed, since it comprises novelty and comprehensibility of the information received, the situation in which the search takes place, and so on. I do not further investigate on this issue, as the point here is that these three components can be used in the classification of different kinds of relevance.

Therefore, a surrogate (a document, some information) is relevant to a query (request, PIN, RIN) with respect to one or more of these components.

The set corresponding to this fourth dimension, the *set of components*, can be defined as:

$$\begin{aligned} Comp \;=\; & \mathcal{P}(Topic, Task, Context) - \emptyset = \\ & \{\{Topic\}, \{Task\}, \{Context\}, \\ & \{Topic, Task\}, \{Topic, Context\}, \{Task, Context\}, \\ & \{Topic, Task, Context\}\} \end{aligned}$$

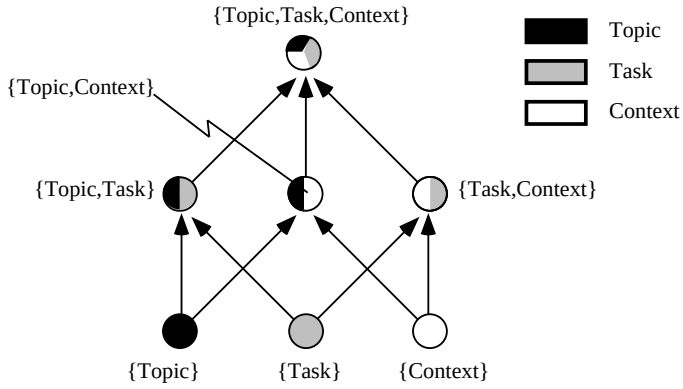

Figure 5: The ordering on *Comp*.

(where $\mathcal{P}(x)$ is the set of all the subsets of x and $\emptyset$ is the empty set). Hence, each of the elements of this set is one or more of the components, in some combination. On this set, an (as said, partial) order can be defined using the set inclusion $\subset$:

$$\forall x, y \in Comp\ (x < y \textbf{ iff } x \subset y).$$

So that, for instance, $\{Topic\} < \{Topic, Task\} < \{Topic, Task, Context\}$, and $\{Task\} < \{Task, Context\} < \{Topic, Task, Context\}$, and so on, as it is graphically represented in Figure 5, using both the arrows and the three colors (black, grey, and white).

2.5 Relevance as a point in a four-dimensional space

Summarising, each relevance can be seen as a point in a four-dimensional space, the values of each dimension being:

1. $InfRes = \{Surrogate, Document, Information\}$;
2. $Repr = \{Query, Request, PIN, RIN\}$;
3. $Time = \{t(rin_0), t(pin_0), t(r_0), t(q_0), t(q_1), t(r_1), t(q_2), \ldots, t(pin_p), \ldots, t(r_m), \ldots, t(q_n), t(f)\}$;
4. $Comp = \{\{Topic\}, \{Task\}, \{Context\}, \{Topic, Task\}, \{Topic, Context\}, \{Task, Context\}, \{Topic, Task, Context\}\}$.

The situation described so far is (partially) represented in Figure 6, that extends Figure 3 with the fourth dimension. This fourth dimension is more difficult to represent, since its elements are only partially (not totally) ordered: in figure, the three colors black, grey and white represent the three components (induced in each of the relevances by the elements of the first two dimensions), emphasising that a relevance may concern one or more of them. The time dimension is not represented.

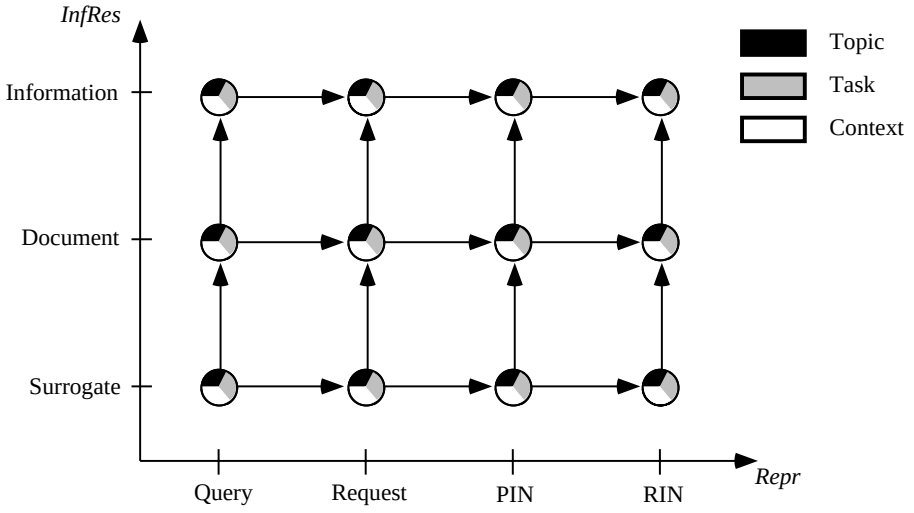

Figure 6: The various kinds of relevance.

More formally, the *partially ordered set of the relevances* can be defined as a cartesian product of the four previous sets:

$$Relevances = InfRes \times Repr \times Time \times Comp.$$

Then, each relevance can be represented by $rel(x, y, t, z)$. For instance:

$$rel(Surrogate, Query, t(q_0), \{Topic\})$$

stands for the relevance of a surrogate to the query at time $t(q_0)$ with respect to the topic component (the relevance judged by an IR system), and

$$rel(Information, RIN, t(f), \{Topic, Task, Context\})$$

stands for the relevance of the information received to the RIN at time $t(f)$ for all the three components (the relevance the user is interested in). In the following I feel free of omitting parameters for indicating a set of relevance kinds when this does not rise ambiguities; for instance, $rel(Information, RIN)$ stands for a not better specified relevance of information to the RIN.

Now, on the set *Relevances* it is possible to formally define the partial order, denoted by $\prec$, in the following way:

$\forall x_1, y_1 \in InfRes,\ \forall x_2, y_2 \in Repr,\ \forall x_3, y_3 \in Time,\ \forall x_4, y_4 \in Comp,$
$rel(x_1, x_2, x_3, x_4) \prec rel(y_1, y_2, y_3, y_4)$
iff
$\forall i (x_i \not\succ y_i) \wedge \exists j (x_j < y_j).$

In Figure 6, the order $\prec$ is graphically represented by the arrows and by the three colors. At this point, the meaning and the usefulness of the four orders defined on the *InfRes*, *Repr*, *Time*, and *Comp* sets should be clear: they allow to

define the $\prec$ order, that represents how much a relevance is near to the relevance interesting for the user, namely

$$rel(Information, RIN, t(f), \{Topic, Task, Context\}).$$

3 Relevance judgement

A relevance judgement is an assignment of a value of *a* relevance by a judge at a certain point of time. Similarly to what done above, it is possible to say that there are many kinds of relevance judgement, that can be classified along five dimensions:

1. The kind of relevance judged (see the previous section);
2. The kind of judge (for instance, it is possible to distinguish between user and non-user);
3. What the judge can use (surrogate, document, or information) for expressing his relevance judgement. It is the same dimension used for relevance, but it is needed since, for instance, the judge can assess the relevance of a document on the basis of a surrogate;
4. What the judge can use (query, request, PIN, or RIN) for expressing his relevance judgement (needed for the same reason of the previous point 3);
5. The time at which the judgement is expressed (at a certain time point, one may obviously judge the relevance in another time point).

This five dimensions can be formalised using five sets, analogously to what done in the previous section (see Mizzaro, 1997a). Note that the two questions "Which (kind of) relevance to judge?" and "Which (kind of) relevance judgement to make?" are different: once decided to judge a given relevance, you can use different relevance judgements for doing that. For instance, it is possible to judge

$$rel(Information, RIN, t(f), \{Topic, Task, Context\})$$

using

$$rel(Surrogate, Query, t(q_0), \{Topic\})$$

(a lower relevance in the $\prec$ order): an IR system does that.

4 Discussion

Some of the consequences of the above proposed framework are analysed in the following subsections: Section 4.1 shows how to classify the types of relevance introduced by other authors, Section 4.2 analyses the design and implementation of more effective IR systems, and Section 4.3 briefly discusses the implications for the evaluation of IR systems.

4.1 The literature on relevance

The classification presented here can obviously be used as a basis for analysing the types of relevance introduced by other authors. For instance (but many other examples could be found, see Mizzaro, 1997a; 1997b):

- Vickery (1959a; 1959b) speaks of 'relevance to a subject' and 'user relevance'. These are $rel(Document, Query, \{Topic\})$ and $rel(Information, RIN, \{Topic, Task, Context\})$, respectively.
- Foskett (1970; 1972) (and many other researchers) distinguishes between 'relevance' ($rel(Request)$) and 'pertinence' ($rel(PIN)$).
- Wilson (1973) explicitly distinguishes $rel(Information, PIN)$ (his 'situational relevance') and $rel(Information, RIN)$.
- Lancaster (1979) defines 'pertinence' as the relation between a document and a request as judged by the user, and 'relevance' as the same relation, but judged by an external judge.
- Swanson (1977; 1986) defines two 'frames of reference' for relevance. Frame of reference one sees relevance as a relation between 'the item retrieved' and the user's need; frame of reference two is based on the user's query. In frame of reference two, relevance is identified with topicality ($rel(Document, Query, \{Topic\})$), and retained more objective, observable and measurable. In frame of reference one, topicality is not enough for assuring relevance ($rel(Document, RIN, \{Topic, Task, Context\})$), a more subjective and elusive notion.
- Soergel (1994) summarises some previously proposed definitions of topical relevance, pertinence and utility: an entity is 'topically relevant' if it can, in principle, help to answer a user's question; an entity is 'pertinent' if topically relevant and 'appropriate' for the user (i.e. the user can understand it and use the information obtained); an entity has 'utility' if pertinent and if it gives to the user 'new' (not already known) information.

Many authors simply distinguish between 'system relevance' and 'user relevance', or they mistake 'system relevance' for 'topic relevance', or they do not consider all the existing kinds of relevance, or they mix relevance and relevance judgements, and so on. It should be clear that some confusion has been made, and that the distinctions proposed in the past are short-sighted if compared with the classification introduced here.

Some other studies compare two different kinds of relevance:

- Cooper (1973a; 1973b) distinguishes between $rel(\{Topic\})$ and $rel(\{Topic, Task, Context\})$;
- Regazzi (1988) compares $rel(Surrogate, Request, \{Topic\})$ and $rel(Surrogate, Request, \{Topic, Task, Context\})$, finding no significant differences;
- Saracevic et al. (1988; 1988a; 1988b) study $rel(Surrogate, Request, \{Topic\})$ and $rel(Surrogate, RIN, \{Topic, Task, Context\})$, called 'relevance' and 'utility', respectively;

- In the experimental evaluation of the FIRE prototype (Brajnik et al., 1996; Mizzaro, 1997a) two different types of relevance are used, namely: $rel(Surrogate, Request, \{Topic\})$ and $rel(Surrogate, Request, \{Topic, Task\})$.

4.2 Towards more effective IR systems

The framework presented in Section 2 can be used to design more effective IR systems for end users, as shown in the following subsections.

4.2.1 Task model

A classical IR system works at the level of $rel(Surrogate, Query, t(q_0), \{Topic\})$; for building IR systems that work at the level of the relevance interesting for the user, namely $rel(Information, RIN, t(f), \{Topic, Task, Context\})$, one has to go up along the order $\prec$ on *Relevances*. For doing that, one can proceed along four independent directions:

- to go up along *InfRes*, for instance using full-text databases instead of bibliographic ones;
- to go up along *Repr*, for instance building IR systems capable of understanding a natural language request;
- to go up along *Time*, implementing systems that allow and stimulate an highly interactive user-system interaction;
- to go up along *Comp*, for instance building beyond topical IR systems, capable of modelling the task and context components.

The first three directions are the classical ones, and many researchers are working on these issues. The fourth direction is particularly interesting: the modelling of beyond topical components of information need has been invoked for many years and by different researchers (Belkin et al., 1982a; Belkin et al., 1982b; Ingwersen, 1992; Barry, 1994; Bruce, 1994), but there is no final answer to this research issue yet. I believe that modelling the context component is difficult, even because it needs further analysis for being completely understood. But the situation is different for the task component, as one could build a system that, starting from some characteristics of the task, infers the desired characteristics of the information (documents, surrogates). Let us analyse more in detail this possibility.

A first idea could be to equip an IR system with a set of stereotypes of tasks, and to allow the user to pick up among them the most suited one. Then the IR system should translate the task into *Concrete Characteristics of Documents* (CCD), maybe going through some sort of characteristics of information, associated with each task stereotype in the set. The problem is that the characteristics of information are dependent not only on the task, but also on the other components (for instance, on the knowledge that the user has on the topic). So, another solution is needed.

A less ambitious, but more feasible, idea is to allow the user to specify some *Abstract Characteristics of Documents* (ACD), and to build a system that can translate the ACD in CCD, thus allowing the user to work at a high level of

abstraction, independent from the particular data base. A possible set of ACD could be:

1. Comprehensibility: how much the retrieved documents have to be easy to understand;
2. Recency: how much the retrieved documents have to be recent;
3. Quantity: how much information the user wants (number of documents and their length);
4. Language: in which language the retrieved documents have to be written;
5. Fertility: how much the retrieved documents will be useful for finding other documents (*eg.* number of references of the documents).

And a possible set of CCD, obtained analysing the INSPEC data base, could be:

- DT (Document type): type of document, for instance BC (Book Chapter), BK (Book), CA (Conference Article), JA (Journal Article), DS (Dissertation), RP (Report), and so on;
- TR (Treatment code): document character, for instance: A (Application), B (Bibliography), G (General or Review), T (Theoretical or Mathematical), and so on;
- PG (Pages): number of pages of the document;
- MD (Meeting date), PD (Publication date) and P5 (Original Publication Date): when the document has appeared;
- LA (Language): language used for writing the document.

One could also take into account the abstract, that contains the number of references of the document and some other hint on how much the document is introductory, theoretical, and so on. The use of a commercial and widely available data base, as INSPEC is, and the generality of the chosen CCD (that can easily be found in other data bases) assure an immediate practical usefulness of this approach.

Let us consider two examples:

1. In Italy, a university professor has to prepare in a few hours an introductory lesson on the UNIX system for a first year course. The professor has a computer science background, but he does not remember a lot about that topic. Moreover, he needs to give a good bibliographic reference to his students. Thus, he will need a few introductory and (preferably) written in Italian documents, and he will not be interested in their recency and fertility.
2. An Italian PhD student has to prepare his PhD thesis on a topic that he does not know very well, and he wants to verify some ideas that he deems promising and original. In this case, comprehensibility is not a crucial aspect, while recency, quantity, and fertility are important. The language of the documents must be Italian or English, the only ones that he knows.

Table 1 summarises the ACD of the retrieved documents. '+' means that the corresponding ACD is important, '–' that it is not important, and '=' that has a medium importance. Note that these ACD derive only from the task and context components of the user's need.

	Professor	PhD student
1. Comprehensibility	High (+)	High (–)
2. Recency	(–)	Recent (+)
3. Quantity	Low (+)	High (+)
4. Language	Italian (=)	Italian or English (+)
5. Fertility	(–)	High (+)

Table 1: ACD of the documents in the two examples.

Now let us see how to map the ACD in CCD. In the first case (professor), we can surely reject documents with 'Document type' CA or DS; the number of pages (PG) must be low, the dates (MD, PD, P5) and the number of bibliographic references are not important and the language (LA) is preferably Italian. In the second case (PhD student), the dates must be recent, the language Italian or English, and the other characteristics are not important.

It seems so feasible to obtain the CCD from the ACD. The obtained CCD could be used in two ways, either for modifying the query (the professor can reject with a high certainty CA and DS documents), or for ranking the retrieved documents (old documents could anyway be interesting for the PhD student).

To have a quantitative idea of the performance improvement that can be obtained by means of an IR system that models the task of the user, let us think of a user interested in documents of a certain kind (for instance, books) and let us suppose that the data base is equally partitioned into three different kinds of documents, for instance books, journal articles, and proceedings articles. If the task component is not taken into account, then probably at least 2/3 of the retrieved documents will be not relevant (though topical); with an IR system capable of modelling the task, the performance (actually, precision) could ideally be three times higher.

4.2.2 Presentation of information

The classification could be useful also for the presentation of information issue. It could be possible to design a user interface to an IR system capable of visualising in some way (*eg.* using virtual reality techniques) the four dimensional relevance space, where the documents could be represented as points (analogously to the 'starfield displays', Ahlberg and Shneiderman, 1994) and directly manipulated by the user. In such an interface, the system could present to the user the most relevant (with respect to each of the relevances) documents, and the user could browse them, read their content, move them in the relevance space, and so on. Such a direct manipulation interface, besides providing the user with an intuitive visualisation mechanism that seems to have good performances, could allow the system to have some feedback from the user. For example, the user could move (dragging with the mouse) a document from the $rel(\{Topic, Task\})$ zone to the $rel(\{Topic\})$ zone, thus allowing the system to

know that some features of that document make it not suited for the task at hand.

4.2.3 Relevance feedback

The classification takes into account the relevance feedback activity: the relevance expressed by the user during the interaction with the IR system is different from $rel(Information, RIN, t(f), \{Topic, Task, Context\})$, hence a document judged relevant before could be not relevant later.

Besides that, the classification could be the basis of a relevance feedback activity richer than the classical one. In classical relevance feedback (Harman, 1992), the user judges the retrieved documents as either relevant or not relevant. But saying that a document is relevant (not relevant) *with respect to a particular relevance* gives additional information that can be fruitfully used by the IR system.[4] For instance, a document already known by the user is not relevant with respect to $rel(\{Context\})$, but it is relevant with respect to $rel(\{Topic, Task\})$. And such a document is obviously a good candidate for a positive feedback, as it contains both topical terms and characteristics suited for the task at hand.

4.3 Evaluation of IR systems

Which is the relevance to use in the IR *evaluation*? The classifications of relevances and relevance judgements could be used as a useful framework on which basis to better understand the evaluation issues. Briefly, in the classical IR system evaluations (Cranfield, see Cleverdon et al., 1966, and TREC, see Harman, 1993), the relevance used, $rel(Surrogate, Request, \{Topic\})$, is a "lower" one in the ordering of Figure 6, but the attempts to climb the ordering (Borlund and Ingwersen, 1997) face many problems. There seems to be a sort of "Relevance indetermination phenomenon" (borrowing the term from quantum physics): the more we try to measure the "real" (user) relevance, the less we can measure it. So the right compromise must be found. Finally, it should be observed that recall and precision are still meaningful and useful aggregates for each kind of relevance.

5 Conclusions and future work

In (Saracevic, 1996) a system of various interplaying relevances is proposed. The classification of various relevances presented in this paper has a narrower range, but it has also some important features:

- It is very schematic and formalised;
- It is useful for avoiding ambiguities on which relevance (and relevance judgement) we are talking about;
- It shows how it is short-sighted to speak merely of 'system relevance' (the relevance as seen by an IR system) as opposed to 'user relevance' (the relevance in which the user is interested), and how 'topicality' (a

[4] This information could be communicated to the IR system by a user using a direct manipulation interface similar to the one described in Section 4.2.2.

relevance for what concerns the topic component) is conceptually different from 'system relevance';

- It has to be considered in the implementation of IR systems working closer to the user, as it is possible to improve a classical IR system along the four independent directions, as above discussed. Moreover, some *postulates of impotence* (borrowing the term from Swanson, 1988) can be straightforwardly derived from the classification, for instance: $rel(Request, t(q_n))$ is the maximum relevance that can be handled with certainty by an IR system; $rel(Surrogate)$ is the maximum relevance that can be handled when using bibliographic databases, and so on;

- It emphasises that it is not so strange that different studies on relevance obtain different results (see Mizzaro, 1997b), as one should pay attention to both which kind of relevance is measured and which kind of relevance judgement is adopted: often this issues are not taken into account.

This work is at an initial stage, and there is still a lot to be done. Possible research questions could be: Are the four dimensions correct? Do we need other dimensions? Is it possible to find an intuitive graphical representation better than the one in Figure 6? Is the decomposition in topic, task, and context correct? Is it possible to extend it, refine it, and define it in a more formal way? Is it possible to improve in some way the orderings on the four dimensions?[5]

Finally, in this paper an ordering has been defined, but it might be interesting to find also a *metric* in order to measure the distances among the various relevances. For doing this, I think it is mandatory to proceed in an experimental way, confronting two or more different kinds of relevance. The four researches briefly summarised at the end of Section 4.1 (Cooper, 1973a; Cooper, 1973b; Regazzi, 1988; Saracevic et al., 1988; Saracevic and Kantor, 1988a; Saracevic and Kantor, 1988b; Brajnik et al., 1996; Mizzaro, 1997a) are some preliminary steps in this direction. This line of research has important practical consequences: with such a metric, it would be possible to choose in an objective way among the possible directions (described in the Section 4.2.1) for developing more effective IR systems.

Acknowledgements

Thanks to Mark Magennis, Giorgio Brajnik, Marion Crehange, Peter Ingwersen, Giuseppe O. Longo, Carlo Tasso, and three anonymous referees for useful discussions and comments on earlier versions of this paper.

References

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

[5] For instance, an alternative could be to use the set cardinality for ordering *Comp*, thus obtaining a more strict ordering ($\{Topic\} < \{Task, Context\}$, and thus $rel(\{Topic\}) \prec rel(\{Task, Context\})$, while these two relevances are not comparable with the order proposed in Section 2.5).

Barry, C. L. (1994). User-defined relevance criteria: An exploratory study, *Journal of the American Society for Information Science* **45**(3): 149–159.

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

Belkin, N. J., Oddy, R. N. and Brooks, H. M. (1982b). ASK for information retrieval: Part II. Results of a design study, *Journal of Documentation* **38**(3): 145–164.

Borlund, P. and Ingwersen, P. (1997). The development of a method for the evaluation of interactive information retrieval systems, *Journal of Documentation* **53**(3): 225–250.

Brajnik, G., Mizzaro, S. and Tasso, C. (1996). Evaluating user interfaces to information retrieval systems: A case study on user support, *SIGIR96, 19th International Conference on Research and Development in Information Retrieval*, Zurich, Switzerland, pp. 128–136.

Bruce, H. W. (1994). A cognitive view of the situational dynamism of user-centered relevance estimation, *Journal of the American Society for Information Science* **45**(3): 142–148.

Cleverdon, C. W., Mills, J. and Keen, M. (1966). *Factors Determining the Performance of Indexing Systems, Vol. 1: Design, Vol. 2: Test results*, College of Aeronautics, Cranfield, UK.

Cooper, W. S. (1971). A definition of relevance for information retrieval, *Information Storage and Retrieval* **7**(1): 19–37.

Cooper, W. S. (1973a). On selecting a measure of retrieval effectiveness, part 1: The "subjective" philosophy of evaluation, *Journal of the American Society for Information Science* **24**(2): 87–100.

Cooper, W. S. (1973b). On selecting a measure of retrieval effectiveness, part 2: Implementation of the philosophy, *Journal of the American Society for Information Science* **24**(6): 413–424.

Cuadra, C. A. and Katter, R. V. (1967a). Experimental studies of relevance judgements, *NSF report TM-3520/001, 002, 003*, Systems Development Corporation, Santa Monica, CA. 3 vols.

Cuadra, C. A. and Katter, R. V. (1967b). Opening the black box of "relevance", *Journal of Documentation* **23**(4): 291–303.

Cuadra, C. A. and Katter, R. V. (1967c). The relevance of relevance assessment, *Proceedings of the American Documentation Institute*, Vol. 4, American Documentation Institute, Washington, DC, pp. 95–99.

Foskett, D. J. (1970). Classification and indexing in the social sciences, *ASLIB Proceedings*, Vol. 22, pp. 90–100.

Foskett, D. J. (1972). A note on the concept of "relevance", *Information Storage and Retrieval* **8**(2): 77–78.

Froehlich, T. J. (1994). 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.

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

Harman, D. (1992). Relevance feedback revisited, *in* N. Belkin, P. Ingwersen and A. M. Pejtersen (eds), *Proceedings of the 15th ACM SIGIR*, ACM Press, New York, pp. 1–10.

Harman, D. (1993). Overview of the first TREC conference, *in* R. Korfhage, E. Rasmussen and P. Willet (eds), *Proceedings of the 16th ACM SIGIR*, ACM Press, New York, pp. 36–47.

Ingwersen, P. (1992). *Information Retrieval Interaction*, Taylor Graham, London.

JAS (1994). Special topic issue: Relevance research, *Journal of the American Society for Information Science* **45**(3).

Lancaster, F. W. (1979). *Information Retrieval Systems: Characteristics, Testing and Evaluation*, 2nd edn, John Wiley and Sons, New York.

Mackay, D. M. (1960). What makes the question, *The Listener* **63**(5): 789–790.

Mizzaro, S. (1995). Le differenti *relevance* in *information retrieval*: una classificazione, *Proceedings of the annual conference AICA'95*, Vol. I, pp. 361–368. In Italian. Translation of the title: "The various relevances in information retrieval: a classification".

Mizzaro, S. (1996a). A cognitive analisys of information retrieval, *in* P. Ingwersen and N. O. Pors (eds), *Information Science: Integration in Perspective — Proceedings of CoLIS2*, The Royal School of Librarianship, Copenhagen, Denmark, pp. 233–250. Paper awarded with the "CoLIS2 Young Scientist Award".

Mizzaro, S. (1996b). How many kinds of relevance in IR?, *in* M. D. Dunlop (ed.), *Proceedings of the Second Mira Workshop*, Monselice, Italy. University of Glasgow Computing Science Research Report TR-1997-2, `http://www.dcs.gla.ac.uk/mira/workshops/padua_procs/`.

Mizzaro, S. (1996c). How many relevances in IR?, *in* C. W. Johnson and M. Dunlop (eds), *Proceedings of the Workshop 'Information Retrieval and Human Computer Interaction', GIST Technical Report GR96-2, Glasgow University*, The British Computer Society, Glasgow, UK, pp. 57–60.

Mizzaro, S. (1997a). *Il Reperimento delle informazioni: analisi teorica e sperimentazione*, Phd thesis in Engineering of Information, University of Trieste (Italy). In Italian. The translation of the title is "Information retrieval: theoretical analysis and experimentation".

Mizzaro, S. (1997b). Relevance: The whole history, *Journal of the American Society for Information Science* **48**(9): 810–832.

Rees, A. M. and Schulz, D. G. (1967). A field experimental approach to the study of relevance assessments in relation to document searching. 2 vols., *NSF Contract No. C-423*, Center for Documentation and Communication Research, School of Library Science, Case Western Reserve University, Cleveland, Ohio.

Regazzi, J. J. (1988). Performance measures for information retrieval systems—An experimental approach, *Journal of the American Society for Information Science* **39**(4): 235–251.

Salton, G. (1989). *Automatic Text Processing – The transformation, Analysis, and Retrieval of Information by Computer*, Addison-Wesley, Reading, MA.

Saracevic, T. (1975). 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.

Saracevic, T. (1996). Relevance reconsidered '96, *in* P. Ingwersen and N. O. Pors (eds), *Information Science: Integration in Perspective — Proceedings of CoLIS2*, The Royal School of Librarianship, Copenhagen, Denmark, pp. 201–218.

Saracevic, T. and Kantor, P. (1988a). A study of information seeking and retrieving. II. Users, questions, and effectiveness, *Journal of the American Society for Information Science* **39**(3): 177–196.

Saracevic, T. and Kantor, P. (1988b). A study of information seeking and retrieving. III. Searchers, searches, and overlap, *Journal of the American Society for Information Science* **39**(3): 197–216.

Saracevic, T., Kantor, P., Chamis, A. and Trivison, D. (1988). A study of information seeking and retrieving. I. Background and methodology, *Journal of the American Society for Information Science* **39**(3): 161–176.

Schamber, L. (1994). Relevance and information behavior, *Annual Review of Information Science and Technology*, Vol. 29, pp. 3–48.

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

Soergel, D. (1994). Indexing and retrieval performance: The logical evidence, *Journal of the American Society for Information Science* **45**(8): 589–599.

Swanson, D. R. (1977). Information retrieval as a trial-and-error process, *Library Quarterly* **47**(2): 128–148.

Swanson, D. R. (1986). Subjective versus objective relevance in bibliographic retrieval systems, *Library Quarterly* **56**: 389–398.

Swanson, D. R. (1988). Historical note: Information retrieval and the future of an illusion, *Journal of the American Society for Information Science* **39**(2): 92–98.

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

van Rijsbergen, C. J. (1979). *Information Retrieval*, 2nd edn, Butterworths, London.

Vickery, B. C. (1959a). The structure of information retrieval systems, *Proceedings of the International Conference on Scientific Information*, Vol. 2, National Academy of Sciences, Washington, DC, pp. 1275–1290.

Vickery, B. C. (1959b). Subject analysis for information retrieval, *Proceedings of the International Conference on Scientific Information*, Vol. 2, National Academy of Sciences, Washington, DC, pp. 855–865.

Wilson, P. (1973). Situational relevance, *Information Storage and Retrieval* **9**(8): 457–471.