Measuring the Agreement Among Relevance Judges

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

The importance of the issue of the agreement (or disagreement) between relevance judges is increasing, since new kinds of relevance judgment expression are being used (to the classical dichotomous one, various researches have added scalar, weighted, and orders of various kind) and new media are being introduced (it is far quicker to judge the relevance of an image than a text, and thus the human judgments can be obtained more easily).

This paper presents a coherent account of the disagreement between relevance judges and groups of judges. Judgment expressions of different kinds, grouped into two categories, are taken into account. To the first category, *score judgments*, belong the more classical dichotomous, scalar, and weighted. To the second one, *order judgments*, belong total (or linear) and partial (or weak) orders, both with or without equality. A uniform notation for describing relevance judgments of each kind is proposed; some of the problems arising when one tries to operationally measure the disagreement between judges are described; a measure for the disagreement of two judges expressing two judgments of the same kind is proposed; the disagreement of a group of more than two judges is discussed; and, finally, some experimental activity inspired by this study is sketched.

Keywords: agreement between relevance judges, disagreement between relevance judges, scalar judgments, order judgments.

1 Introduction

In the Information Retrieval (IR) field, the expression "agreement among judges" (or "disagreement among judges") is used quite often when investigating and using relevance judgments. Various researchers have studied the disagreement between two judges, between the judges in a group, between two groups of judges, and even between one judge and herself, in different periods of time (this is named *consistency* of a judge).

At a first glance, this might seem a not interesting issue: two or more judges agree if all of them judge relevant, or not relevant, a document; otherwise they disagree. But if it is easy to say what does it mean that some judges completely agree, it is not obvious what does it mean that they completely disagree, and it is even less obvious how to give a *measure* of the disagreement. And this situation becomes even more complex if one takes into account various kinds of relevance judgment expression (dichotomous, using a rating scale, continuous, order of documents, and so on) and various forms of agreement (between two judges, between N judges, between groups of judges, and so on).

Such a measure of judges agreement would be important for many reasons:

- To empirically study how the kind of expression of the judgment affects the agreement. It could be, for instance, that the agreement results higher, or lower, when using dichotomous judgments than when using order judgments. Then, once understood how the kind of expression affects the agreement, one can compare the agreement between groups of judges using different kinds of judgment expression.

- To compare the agreement of the judges among different test collections, even if the relevance judgments are expressed in different ways. If different values for different test collections are found, it might be worthwhile to study the reasons and to try to increase the agreement, for instance, introducing other non-topical components, as suggested in [5]. With research activities like this one, it should be possible to find a confirmation, or a refutation, of the quite often maintained claim that variations between relevance judges are (even) greater in a multimedia context than a textual one. One could also plan to design test collections with different degrees of agreement, to be used with different purposes: a high agreement is not necessarily a positive feature, if it causes a not realistic situation.
- To better understand the experiments aimed at evaluating IR systems with judgments by real users. A low agreement among the users judging the relevance of the retrieved documents might be a clue of an inadequate presentation of the documents, or of a lack of care by the subject, or of a low reliability of the results.
- To deal with new kinds of relevance judgment expression (to the classical dichotomous one, various researches have added scalar, weighted, and orders of various kind) and with new media (it is far quicker to judge the relevance of an image than a text, and thus human judgments can be obtained more easily).

In this paper I try to clarify this issue, defining some operational measures of judges (dis)agreement for different kinds of relevance judges expression.

This paper is organized as follows: Section 2 describes some previous work on this issue; Section 3 proposes a classification of the various kinds of relevance judgments; Section 4 provides a uniform notation for relevance judgments of various kinds; Section 5 contains some examples showing which problems could arise when trying to measure the disagreement between judges; Section 6 proposes some ways for measuring the disagreement between two judges using the same kind of judgment expression; Section 7 discusses the issue of the disagreement in a group of more than two judges; Section 8 concludes the paper and sketches some future activity.

2 Related work

Barhydt [1, 2] introduced the following measures for the similarity between users' and non-users' relevance judgments: *sensitivity* (among the documents judged relevant by the user, the percentage judged relevant also by the non-user) and *specificity* (among the documents judged non-relevant by the user, the percentage judged non-relevant also by the non-user). *Effectiveness* is the synthesis of these two measures into a single one:

$$\textit{effectiveness} = \textit{sensitivity} + \textit{specificity} - 1.$$

Moreover, the author compared the (dichotomous) relevance judgments by subject experts and IR experts, finding a 0.35 average effectiveness.

Hoffman [7] studied the consistency of relevance judgments among different groups of judges and among judges of the same group.

O'Connor [13] studied the effects of unclear requests on the relevance judgment: if the request is unclear then different judges will interpret it differently, and the agreement among them will be low. He also suggested, on the basis of experimental evidence, that a discussion among judges modifies relevance judgments and can resolve disagreements [14].

Lesk and Salton [11] found a 30% agreement between users' and non-users' relevance judgments. They defined a *strong hypothesis* (difference in relevance judgments cannot affect the assessment of retrieval performance) and a *weak hypothesis* (differences in relevance judgments cannot affect the comparison of performances of different retrieval methods). Both hypotheses were supported by experimental data.

Figueiredo [4] found a 57.2% agreement between librarians' and users' relevance judgments on a three-point scale.

Kazhdan [10] found experimental evidence to support the weak hypothesis of Lesk and Salton, but not to support their strong hypotheses.

Rorvig [15] proposed to substitute the usual relevance judgments with "preference" judgments, *i.e.*, judgments of preference of one document over another one. He showed some experimental results that seem to confirm the reliability of this approach.

Burgin [3] found good agreements (from 40% to 55%) among judges of 4 different groups (users, online searching experts, and two kinds of subject experts, "more" and "less" expert) judging full-text documents.

Janes and McKinney [9] compared users' relevance judgments with non-users' (students of information/library studies and psychology), finding a 0.62 specificity and a 0.68 sensitivity.

Janes [8] compared users' relevance judgment with non-users' judgment of: relevance (not defined), topicality (similarity to the topic), and utility (usefulness to the user). The judges belonged to three different groups: incoming students to a school of information/library science, experienced students in that school and academic librarians. The study is an exploratory one, and the results are synthesized in the following table, adapted from [8]:

	Sensitivity	Specificity
Incoming students	0.861	0.557
Experienced students	0.778	0.844
Library staff	0.694	0.773

Harter [6] analyzed the literature concerning the factors affecting the relevance judgments and the experimental evaluation of IR systems. He derived that Lesk and Salton's weak hypothesis, *i.e.*, the assumption that the variations in relevance judgments do not significantly affect the measurement of IR systems performance (on which the Cranfield-like experiments are based), is not supported. He suggested a new approach to evaluation experiments, in which different "problem types" (different types of searchers, request, and relevant documents) are evaluated separately.

Another related study, even if it not deals directly with relevance judgments, is [16], that defined a measure of the distance between two orderings with the purpose of measuring the performance of a retrieval system.

3 Different kinds of judgment expression

Given a set of documents, the following kinds of relevance judgments can be adopted:

- **Score judgments.** The judge assigns a value, or score, to each document. There are three sub-cases:
  - *Dichotomous*: the judge assigns a relevant/not relevant (yes/no, 0/1) judgment to each document in the set. This is a particular case of the following.
  - *Scalar*: the judge assigns a value taken from a scale, usually of 3–7 values (for instance, not relevant, partially relevant, relevant), to each document in the set. This is a particular case of the following.
  - *Weighted*: the judge assigns a value that is a real number, usually in the [0–1] range, to each document in the set.
- **Order judgments.** The judge does not assign a value, but orders the documents. There are two sub-cases:
  - *Total orders*: each document has to be comparable to the other ones: in mathematical terms, the ordering must be *total*, or *linear* (see Figure 1a). Given two documents, the judge cannot say that she does not know how to compare them. Again, we have two sub-cases:
    - Without equality: the ordering relation is '$<$'; no document can be equally relevant to another document.
    - With equality: the ordering relations are '$<$' and '$=$'; some documents can be equally relevant to other documents.[1]
  - *Partial orders*: it is not necessary that each document is comparable to the other ones: the order can be a *partial* (or *weak*) one (see Figure 1b: neither b, c, and d nor b and e are comparable). Given two documents, the judge is allowed to say that she does not know how to relate them. Note that this is not a purely hypothetical possibility: if one adopts a *multidimensional* relevance [5, 12] this is the normal case. We have the same two sub-cases as before: without and with equality.

[1] The distinction without/with equality might be done even in score judgments, but in that case it is not an interesting distinction; we will see in the following that it is instead important for the order judgments.

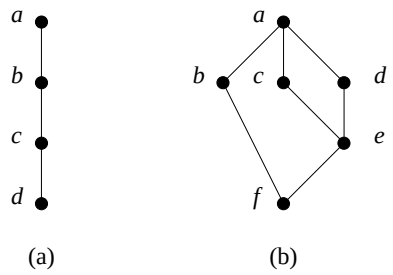

Figure 1: Graphical representation of total (a) and partial (b) orders.

Given a score judgment, it is always possible to derive a total order with equality. Sometimes, since the weight is not used in a consistent way, the order is the only reliable data that can be extracted from the weighted judgments of a set of documents.

4 Notation

In the following, small capital letters $a, b, c, \ldots$ stand for documents. Instead of speaking of disagreement among judges, I will speak of disagreement among *judgments*, and I will indicate the judgments with $j_1, j_2, \ldots, j_i, \ldots$. One judgment does not necessarily concern one document only (because it would make no sense to order one document). The number of documents judged is said the *cardinality* of the judgment, and is abbreviated with *card*(j_i). I will only speak of disagreement between judgments with the same cardinality.

I will use square brackets for judgments of every kind: for the score judgments, the square brackets contain the score assigned to each document, separated by commas; for the ordering judgments, the square brackets contain the documents and the relations among them (or a list of such relations separated by commas). The relations used are: '$<$' for "less relevant", '$=$' for "equally relevant", and '?' for "not comparable". I will use the parentheses for avoiding ambiguities. I will sometimes omit the '$<$' in the total order without equality, when this does not rise any ambiguity.

Thus, for instance, for three documents $\{a, b, c\}$, a dichotomous judgment could be $j_1 = [R, N, R]$ (a and c are relevant, b is not relevant); a three-point scale scalar judgment could be $j_2 = [R, N, P]$ (a is relevant, b is not relevant, c is partially relevant); a weighted [0–1] judgment could be $j_3 = [0.9, 0.2, 0.7]$; a total order without equality judgment could be $j_4 = [b{<}c, c{<}a, b{<}a]$ or, more shortly, $[b{<}c{<}a]$, or even $[bca]$ (b is less relevant than c that is less relevant than a); a total order with equality could be $j_5 = [b{<}a, a{=}c]$ or $[b{<}a{=}c]$ (a is equally relevant than c and both of them are more relevant than b); a partial order without equality could be $j_6 = [b{<}(a?c)]$ (the judge recognizes that both a and c are more relevant than b, but she cannot say if a is less, more, or equally relevant than c; note that the parentheses are needed since $[(b{<}a)?c]$ would be a different judgment); and a partial order with equality could be $j_7 = [(a{=}b)?c]$ (a and b are equally relevant, but nothing is said on their relation with c). All the j_1–j_7 judgments have the same cardinality, 3.

If j_i is a score judgment, by $j_i(k)$ I will refer to the score given to the k-th document of the set being judged; for instance, using the above presented judgments, $j_1(2) = N$, $j_3(3) = 0.7$, and so on.

I will indicate as $d(j_1, j_2)$ the disagreement between the two judgments j_1 and j_2 (that must be of the same kind). It will be a real number in the [0–1] range.

5 Examples and problems

To give a more concrete description and to show which problems could arise, let us take a set of five documents $\{a, b, c, d, e\}$ and three judges expressing three judgments j_1, j_2, j_3. We could have the following typical cases and problems/questions:

- **Dichotomous.** The three judges might express the following three judgments: $j_1 = [R, R, R, R, R]$, $j_2 = [R, R, R, N, N]$, and $j_3 = [N, N, N, N, N]$ (where R stands for relevant and N for not relevant). Of course, j_1 and j_3 have the maximum disagreement.

- **Scalar.** With a three values scale, we could have $j_1 = [N, R, N, R, R]$, $j_2 = [R, N, R, N, R]$, and $j_3 = [P, P, P, P, P]$. Here it is less obvious to decide which judges disagree to a greater extent: j_1 and j_2 completely disagree on the first four documents, but they agree on the fifth, while j_3, even if uses completely different values from both j_1 and j_2 for all the five documents, seems to have a higher agreement with both of them.

- **Weighted.** The three judges might express the following judgments: $j_1 = [0.1, 0.9, 0.1, 0.9, 0.9]$, $j_2 = [0.9, 0.1, 0.0, 0.1, 0.9]$, and $j_3 = [0.5, 0.5, 0.5, 0.5, 0.5]$. We have a situation similar to the scalar case: j_1 and j_3 have quite different scores for the first four documents, and the same score for the fifth one. Judgment j_2 has different values than j_1 and j_3, but the differences are lower. Of course the maximum disagreement is between $[0, 0, 0, 0, 0]$ and $[1, 1, 1, 1, 1]$, or between $[0, 1, 0, 1, 0]$ and $[1, 0, 1, 0, 1]$. The previous two kinds of judgment (dichotomous and scalar) can be seen as weighted ones, for instance with $N = 0$, $P = 0.5$, and $R = 1$.

- **Total order without equality.** The three judges might express the following judgments: $j_1 = [a<b<c<d<e]$, $j_2 = [e<d<c<b<a]$, and $j_3 = [e<a<b<c<d]$. We could assume that j_1 and j_2 have the maximum disagreement since they are opposite orders; but we could also maintain that j_1 and j_3 have a higher disagreement because j_1 and j_2 have one document (c) in the same position. It is not obvious how to measure the disagreement: we could use the number of different positions in two orders, but this could be not adequate (j_1 and j_2 would disagree less than j_1 and j_3).

- **Total order with equality.** The three judges might express the following judgments: $j_1 = [a=b<c<d<e]$, $j_2 = [a=b=c<d<e]$, and $j_3 = [a<b<c<d<e]$. Which judge (j_2 or j_3) does disagree more with j_1? How to extend the previous case (total order without equality) in a consistent way?

- **Partial order without equality.** The three judges might express the following judgments: $j_1 = [(a?b)<c<d<e]$, $j_2 = [(a?b?c)<d<e]$, and $j_3 = [a<b<c<d<e]$. Again, which judge does disagree more with j_1? How to extend the total order without equality in a consistent way? And which is the difference between '=' and '?', *i.e.,* how do total with equality and partial without equality orders relate?

- **Partial order with equality.** The three judges might express the following judgments: $j_1 = [(a?b)<c=d=e]$, $j_2 = [(a?b?c)<d<e]$, and $j_3 = [a<b<c<d<e]$. The questions are similar to the ones in the previous cases: which judgments do disagree more? How to extend the total with equality order and the partial order without equality? Note that $[(a?b)=c]$ would not be a normal judgment, since it implies that $[a=b=c]$.

6 Measures of the disagreement between two judges

I propose some measures of the disagreement between two judges, a value in the [0–1] range, calculated in different ways depending on the kind of judgment expression.

6.1 Score judgments

Dichotomous judgment is a particular case of the scalar one, which is a particular case of the weighted one, but I prefer to present them separately since most of the literature studying relevance judgments (for instance, all the papers cited in Section 2 but [15]) deals with dichotomous and scalar judgments.

6.1.1 Dichotomous judgments

The classical dichotomous judgment is the simpler kind of judgment. We can easily define the disagreement between two judges as the frequency of the different judgments; for instance, if a first judgment is $j_1 = [R, R, R, N, N]$ and a

second judgments is $j_2 = [R, R, N, N, R]$, the disagreement is $d(j_1, j_2) = \frac{0+0+1+0+1}{5} = \frac{2}{5}$.[2]

6.1.2 Scalar judgments

Given a scale with n values, we can: define a function v that assigns the values from 0 to $n-1$ to the n values, from the least relevant to the most relevant; define a distance between two judgments as the sum of the absolute differences; and calculate $d(j_1, j_2)$ as the normalized distance between each score in the two judgments. Formally:

$$d(j_1, j_2) = \frac{\sum_{i=1}^{card(j_1)} \frac{|v(j_1(i)) - v(j_2(i))|}{n-1}}{card(j_1)}.$$

For instance, given a five values scale (N for not relevant, L for low relevant, P for partially relevant, H for high relevant, and R for relevant) and the two judgments $j_1 = [R, H, R, N, L]$ and $j_2 = [H, R, L, N, R]$ we would have:

$$d(j_1, j_2) = \frac{\frac{1}{4} + \frac{1}{4} + \frac{3}{4} + \frac{0}{4} + \frac{3}{4}}{5} = \frac{2}{5}.$$

Note that it is not by chance that we obtain the same disagreement obtained for the dichotomous judgment. The judgments j_1 and j_2 in the scalar case are coherent with those in the dichotomous case: in the scalar judgments, the Rs of the dichotomous judgments are changed into Hs or Rs, and the dichotomous Ns into Ns or Ls.

6.1.3 Weighted judgments

In an analogous way to dichotomous and scalar judgments, we can calculate the disagreement between two judgments with the following formula

$$d(j_1, j_2) = \frac{\sum_{i=1}^{card(j_1)} |j_1(i) - j_2(i)|}{card(j_1)}$$

(we do not need anymore neither the v function, nor to normalize, since the single values of each judgment are already in the [0–1] range). For instance, with $j_1 = [0.1, 1, 0.1, 0.9, 0.9]$ and $j_2 = [0.9, 0.1, 0.0, 0.2, 0.9]$ we have

$$d(j_1, j_2) = \frac{0.8 + 0.9 + 0.1 + 0.7 + 0}{5} = 0.5.$$

6.2 Order judgments

The basic idea is that an order is a permutation of some elements, and that we have to calculate the *distance* between two permutations. The normalized distance will be the measure of the disagreement.

6.2.1 Total order judgments

I will start presenting the disagreement for the total order without equality judgments, that is simpler; the order with equality case follows and extends it in a consistent way.

Order without equality. For measuring the distance δ between two orders without equality I propose to use *the minimum number of switches of adjacent elements*. For instance, if we have five documents a, b, c, d, and e, the distance between $j_1 = [abcde]$ and $j_2 = [abced]$ is $\delta(j_1, j_2) = 1$; between j_1 and $j_3 = [abedc]$ is $\delta(j_1, j_3) = 3$, because $[abcde] \rightarrow [abced] \rightarrow [abecd] \rightarrow [abedc]$ (the arrow $\rightarrow$ represents one switch); and between j_1 and $j_4 = [edcba]$ is $\delta(j_1, j_4) = 10$, and this is the maximum distance between two orders (corresponding to the maximum disagreement between two judges) of 5 documents.

[2] Of course, one could also define the *agreement* $a(j_1, j_2) = 1 - d(j_1, j_2)$: in this case the agreement would be 3/5. For avoiding confusion, I will speak of disagreement only.

The maximum distance for n documents is defined recursively as

$$\begin{aligned}\delta^*(1) &= 0\\ \delta^*(n) &= (n-1) + \delta^*(n-1)\end{aligned}$$

since:

- the more distant orders are the opposite ones, for instance $[123\ldots n]$ and $[n\ldots 321]$;
- the $(n-1)$ part of the formula is the number of switches needed to go from $[123\ldots n]$ to $[23\ldots n1]$;
- the $\delta^*(n-1)$ is the number of switches for going from $[23\ldots n1]$ to $[n\ldots 321]$.

And it is easy to see that

$$\delta^*(n) = \sum_{i=1}^{n-1} i = \frac{n\cdot(n-1)}{2}. \tag{1}$$

The number of switches is not a value in the [0–1] range. For obtaining such a value as a measure of the disagreement, we divide the distance δ by the maximum distance δ^*:

$$d(j_1, j_2) = \frac{\delta(j_1, j_2)}{\delta^*(card(j_1))}. \tag{2}$$

Order with equality. If the equality is allowed, the situation is similar to the previous case, but more complex, since we have to take into account both the '$<$' and '$=$' relations. I will gradually define a distance δ, that will later be normalized as done above.

Let us start by observing that we have two possible kinds of *changes*: the switch $[a{<}b] \rightarrow [b{<}a]$; and a new kind of change, when we modify an '$=$' into a '$<$' or vice versa ('$<$' into '$=$'), like in $[a{<}b] \rightarrow [a{=}b]$. The second kind of change seems to lead to a smaller disagreement; in other words, it is reasonable that

$$d([a{<}b], [a{=}b]) < d([a{<}b], [b{<}a]).$$

For obtaining this behavior, I define two different weights for the two kinds of change:

- 1 for the $[a{<}b] \rightarrow [b{<}a]$ switch, as before;
- 0.5 for the $[a{<}b] \rightarrow [a{=}b]$ and $[a{=}b] \rightarrow [a{<}b]$ change.

In this way, we extend in a consistent way the without equality case, since we do not introduce any shortest distance between $[a{<}b]$ and $[b{<}a]$:

$$\begin{aligned}&\delta([a{<}b], [b{<}a]) =\\ &\quad \delta([a{<}b], [a{=}b]) + \delta([a{=}b], [b{=}a]) + \delta([b{=}a], [b{<}a]) =\\ &\quad 0.5 + 0 + 0.5 = 1.\end{aligned}$$

We can calculate for instance

$$\begin{aligned}&\delta([a{=}b{<}c{=}d{<}e], [a{<}b{<}c{<}d{<}e]) =\\ &\quad \delta([a{=}b{<}c{=}d{<}e], [a{<}b{<}c{=}d{<}e]) + \delta([a{<}b{<}c{=}d{<}e], [a{<}b{<}c{<}d{<}e]) =\\ &\quad 0.5 + 0.5 = 1.\end{aligned}$$

But we have a problem. We have seen that

$$\delta([a{<}b{<}c{<}d{<}e], [e{<}d{<}c{<}b{<}a]) = 10,$$

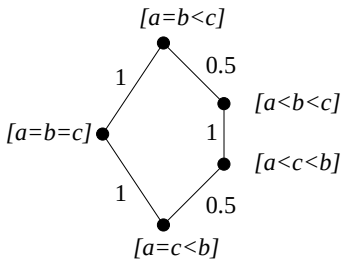

Figure 2: The distance between $j_1 = [a{=}b{<}c]$ and $j_2 = [a{=}c{<}b]$.

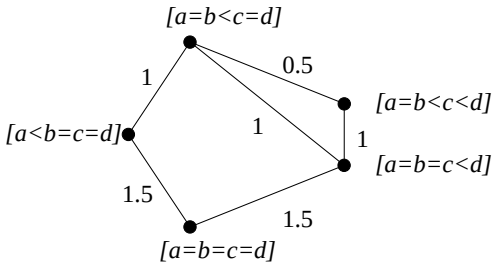

Figure 3: The distance between $j_1 = [a{=}b{<}c{=}d]$ and $j_2 = [a{=}b{=}c{=}d]$.

while

$$\begin{aligned}\delta([a{<}b{<}c{<}d{<}e],[a{=}b{=}c{=}d{=}e]) &= 0.5 \cdot 4 = 2\\ \delta([a{=}b{=}c{=}d{=}e],[e{=}d{=}c{=}b{=}a]) &= 0\\ \delta([e{=}d{=}c{=}b{=}a],[e{<}d{<}c{<}b{<}a]) &= 2.\end{aligned}$$

It seems that we have a shortest distance, since 10 is obviously greater than $2+2$. To avoid this problem, and to obtain a definition that extends consistently the without equality case, some other details have to be added:

- A change from '$<$' to '$=$' has to be multiplied by the length of the '$=$' chain created (measured as the number of '$=$' signs in the chain). For instance, $\delta([a{=}b{<}c],[a{=}b{=}c]) = 0.5 \cdot 2 = 1$. This can also be seen in the following way: $[a{=}b{<}c] \rightarrow [a{<}b{=}c] \rightarrow [a{=}b{=}c]$ (two 0.5 changes).
- A change from '$=$' to '$<$' is calculated as the inverse change from '$<$' to '$=$'.

This avoids the previous problem, but some other cases have to be considered: which is the distance between $j_1 = [a{=}b{<}c]$ and $j_2 = [a{=}c{<}b]$? The change from j_1 to j_2 is not a simple switch between b and c since the '$=$' relation is affected too (and thus $\delta(j_1, j_2) \neq 1$). Actually, $\delta(j_1, j_2) = 2$, as one can see from Figure 2.

As another example, the distance between $j_1 = [a{=}b{<}c{=}d]$ and $j_2 = [a{=}b{=}c{=}d]$ can be calculated in various ways, as shown in Figure 3. Note that it is quite different from $\delta([a{=}b{<}c{<}d],[a{=}b{=}c{<}d]) = 1$: bringing c on the left of the '$<$' sign is a simple switch in the second case, while it "breaks" an equality in the second case.

Once defined the distance δ, the disagreement for the order with equality judgment is defined like in the order without equality case, see Formulae (1) and (2): it is easy to see that the maximum distance δ^* is the same as in the two cases, since the most distant orders are the opposite ones and the '$=$' relations lead to a lower distance than '$<$'.

6.2.2 Partial order judgments

One might argue that the partial order without equality seems similar to the total order with equality, but this is a misleading approach. Another approach is that a partial order (both without and with equality) judgment is a *set*

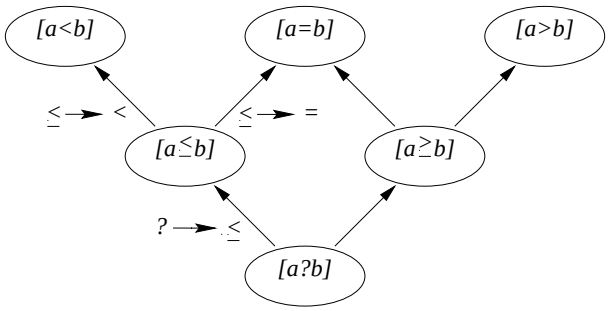

Figure 4: The set of all the orders on the two elements set $\{a, b\}$.

of total order judgments (representing, in this way, uncertainty or lack of knowledge), *e.g.*, $j_1 = [a?b]$ is the set of judgments $\{[a<b], [a=b], [b<a]\}$. This does not lead (at least not in an immediate and natural way) to defining a unique distance between two partial order judgments, but only to *minimal* (δ_{min}) and *maximal* (δ_{max}) distances. For instance, $j_1 = [a?b]$ and $j_2 = [a<b]$ have:

$$\begin{aligned} \delta_{min}([a?b], [a<b]) &= \delta([a<b], [a<b]) = 0 \\ \delta_{max}([a?b], [a<b]) &= \delta([b<a], [a<b]) = 1. \end{aligned}$$

Note that this definition is a consistent extension of the total order case, in which δ_{min} and δ_{max} always have the same value. Obviously, with this definition, we can have only *minimal* and *maximal* disagreements (d_{min} and d_{max}), which can be defined as usual, see Formulae (2) and (1).

More satisfying definitions can be found adopting a more general approach. Given a set S of elements (the documents being judged), let us leave the judge free to express the judgments on S using any of the relations in the set $\Phi = \{?, =, <, \leq, >, \geq\}$. An ordering on S is a set of terns $\langle x, \phi, y\rangle$ where $x, y \in S$ and $\phi \in \Phi$. For instance, if $S = \{a, b, c\}$, two different orderings could be $\omega_1 = [a?b, a<c, b<c]$ and $\omega_2 = [a<b, a<c, b?c]$.

Now we can define an *order space* $\Omega(S)$ as the set of all the possible orderings on S. For defining a distance function on $\Omega(S)$, we can introduce on this set a reflexive, antisymmetric, and transitive accessibility relation (*à la* Kripke), that represents when one order can be derived (adding knowledge in a consistent way) starting from another one. For instance, Figure 4 represents $\Omega(\{a, b\})$ and the accessibility relation on it (the nodes are the possible orderings and the arcs represent the accessibility relation): from $[a?b]$ it is possible to access all the other orders; from $[a\leq b]$ it is possible to access both $[a<b]$ and $[a=b]$, and so on. This accessibility relation can be used for defining a distance between two different orderings as the minimum number of arcs between the two orders, normalized in the usual way (dividing it by the maximum distance in the order space).

Coming back to the example in Figure 4, the maximum distance is between $[a<b]$ and $[a>b]$: 4 arcs that, once normalized, give a distance, between these two orderings, of 1; the distance between $[a?b]$ and $[a\leq b]$ (or $[a\geq b]$) is $1/4$, and the distance between $[a=b]$ and $[a<b]$ (or $[a>b]$) is $1/2$, consistently with the definitions in Section 6.2.1.

The accessibility relation on $\Omega(S)$ can be labeled with *minimal rewriting rules* (having the pattern $\phi \rightarrow \phi'$, where $\phi, \phi' \in \Phi$), that define how to modify an ordering ω_1 (containing the tern $\langle x, \phi, y\rangle$) into another one ω_2 (in which the original tern is changed into $\langle x, \phi', y\rangle$) directly accessible from ω_1. These rules are: $? \rightarrow \leq, ? \rightarrow \geq, \leq \rightarrow <, \geq \rightarrow >$, $\leq \rightarrow =$, and $\geq \rightarrow =$. Some of the arcs in Figure 4 are labeled with the corresponding rewriting rule.

Finally, let us remark that the cardinality of $\Omega(S)$ rapidly increases with the cardinality of S: its value is given by

$$|\Omega(S)| = |\Phi|^{\frac{|S|(|S|-1)}{2}}$$

(even if some of these orderings are not consistent and should be discarded, or dealt with in an appropriate way, as suggested in Section 8). Obviously, this makes difficult to graphically represent $\Omega(S)$ when $|S| > 2$.

J_a	J_b	J_c	J_d	J_e
$a\ b\ c\ d$	$a\ b\ c\ d$	$a\ b\ c\ d$		
$j_1 = [\ 0, 0, 0, 0\ ]$	$j_1 = [\ 0, 0, 0, 0\ ]$	$j_1 = [\ 0, 0, 0, 0\ ]$	$j_1 = [\ a, b, c, d\ ]$	$j_1 = [\ a, b, c, d\ ]$
$j_2 = [\ 0, 0, 0, 0\ ]$	$j_2 = [\ \frac{1}{3}, \frac{1}{3}, \frac{1}{3}, \frac{1}{3}\ ]$	$j_2 = [\ 1, 1, 0, 0\ ]$	$j_2 = [\ a, b, c, d\ ]$	$j_2 = [\ c, d, a, b\ ]$
$j_3 = [\ 1, 1, 1, 1\ ]$	$j_3 = [\ \frac{2}{3}, \frac{2}{3}, \frac{2}{3}, \frac{2}{3}\ ]$	$j_3 = [\ 0, 0, 1, 1\ ]$	$j_3 = [\ d, c, b, a\ ]$	$j_3 = [\ b, a, d, c\ ]$
$j_4 = [\ 1, 1, 1, 1\ ]$	$j_4 = [\ 1, 1, 1, 1\ ]$	$j_4 = [\ 1, 1, 1, 1\ ]$	$j_4 = [\ d, c, b, a\ ]$	$j_4 = [\ d, c, b, a\ ]$

Figure 5: Five groups of judgments.

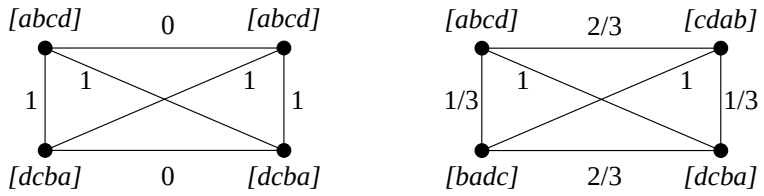

Figure 6: The disagreements between judgments in the J_d and J_e groups.

7 Measures of the disagreement of a group of judges

Until now I have spoken of the disagreement between *two* judgments only. If we have a group $J = \{j_1, j_2, \ldots j_n\}$ of more than two judgments, the *disagreement of the group* can be defined as the average of the disagreement of each judgment with the other ones:

$$D(J) = \frac{\sum_{i=1}^{n} \frac{\sum_{k \neq i} d(j_i, j_k)}{(n-1)}}{n} \tag{3}$$

Let us see some examples. Figure 5 shows five possible sets of judgments (J_a, J_b, J_c, J_d, and J_e) expressed by four judges j_1, j_2, j_3, j_4 on four documents a, b, c, d. J_a, J_b, and J_c contain weighted judgments and J_d and J_e contain total order without equality judgments. We have that (see also Figure 6):

$$D(J_a) = \frac{\frac{0+1+1}{3} + \frac{0+1+1}{3} + \frac{0+1+1}{3} + \frac{0+1+1}{3}}{4} = \frac{\frac{2}{3} + \frac{2}{3} + \frac{2}{3} + \frac{2}{3}}{4} = \frac{2}{3}$$

$$D(J_b) = \frac{\frac{\frac{1}{3}+\frac{2}{3}+1}{3} + \frac{\frac{1}{3}+\frac{1}{3}+\frac{2}{3}}{3} + \frac{\frac{1}{3}+\frac{1}{3}+\frac{2}{3}}{3} + \frac{\frac{1}{3}+\frac{2}{3}+1}{3}}{4} = \frac{\frac{2}{3} + \frac{4}{9} + \frac{4}{9} + \frac{2}{3}}{4} = \frac{5}{9}$$

$$D(J_c) = \frac{\frac{\frac{1}{2}+\frac{1}{2}+1}{3} + \frac{\frac{1}{2}+1+\frac{1}{2}}{3} + \frac{\frac{1}{2}+1+\frac{1}{2}}{3} + \frac{1+\frac{1}{2}+\frac{1}{2}}{3}}{4} = \frac{\frac{2}{3} + \frac{2}{3} + \frac{2}{3} + \frac{2}{3}}{4} = \frac{2}{3}$$

$$D(J_d) = \frac{\frac{0+1+1}{3} + \frac{0+1+1}{3} + \frac{0+1+1}{3} + \frac{0+1+1}{3}}{4} = \frac{\frac{2}{3} + \frac{2}{3} + \frac{2}{3} + \frac{2}{3}}{4} = \frac{2}{3}$$

$$D(J_e) = \frac{\frac{\frac{2}{3}+\frac{1}{3}+1}{3} + \frac{\frac{2}{3}+\frac{1}{3}+1}{3} + \frac{\frac{2}{3}+\frac{1}{3}+1}{3} + \frac{\frac{2}{3}+\frac{1}{3}+1}{3}}{4} = \frac{\frac{2}{3} + \frac{2}{3} + \frac{2}{3} + \frac{2}{3}}{4} = \frac{2}{3}.$$

By comparing $D(J_a)$ and $D(J_b)$, one realizes that with this definition of the disagreement of a group, the maximum disagreement is obtained when half of the judges all give a judgment j, the other half all give a judgment j', and the

(a) (b)

Figure 7: Maximum disagreement in a group of judges.

disagreement between j and j' is the maximum, *i.e.*, $d(j, j') = 1$. We can represent this situation graphically as in Figure 7a.

One might argue that the maximum disagreement of a group of judges is obtained when each judge gives different judgments from the other judges (Figure 7b), since this would correspond to a maximum entropy situation. This is partially true, since J_c and J_e are groups of judgments with the maximum disagreement, but not necessary, as J_a and J_d prove.

With the above proposed definition, the maximum disagreement of a group depends on the cardinality of the group: for a group of n judgments we have

$$D^* = \frac{n}{2 \cdot (n-1)}$$

(since, in the maximum disagreement situation of Figure 7a, the disagreement of each judge with $n/2$ judges is 1, and with the other $n-1$ judges is 0, and this leads to an average disagreement for each judge of $\frac{n/2}{n-1}$). Note that the disagreement of the group is a value in the [0–1] range, but is not normalized: the value 1 is never obtained. Anyway, it would be easy to redefine the disagreement of a group as the one proposed in Formula (3) divided by D^*.

8 Conclusions and future work

This paper has presented a coherent account of the disagreement between relevance judges and groups of judges, for different kinds of judgment expression (both score judgments—dichotomous, scalar, and weighted— and order judgments—total and partial orders, without and with equality). A measure for the disagreement of two judges expressing two judgments of the same kind has been proposed; and the disagreement of a group of more than two judges has been discussed.

There are some mandatory improvements on the theoretical side: multidimensional relevance judgments (as intended in [12, 5]) should be added to the classification proposed in Section 3 and fully taken into account (it is true that a multidimensional relevance judgment can be transformed into a partial order, but, in this way, one loses some information). Moreover, in this paper I have assumed that the judges express their orderings on the whole set of documents. However, if the set is too large, this is not feasible, and another approach is needed. For instance, the judges can be presented, and asked to rank, pairs of documents only: in this way, the obtained ordering might be a particular one (*e.g.*, circular, or incomplete). This issue requires further analysis, together with the related problem of the transitivity of an order: the *transitive closure* can lead to some asymmetry, as the subset of $\Omega(\{a, b, c\})$ in Figure 8 shows.

In the future, besides working on these theoretical extensions, I also plan to perform some experiments for comparing the measures between different kinds of judgment. For instance, different kinds of judgments on the relevance of some documents to some queries could be collected and compared. Also some data from previous experiments could be used; for instance, the data from two experiments performed at Dublin MIRA workshop (http://www.dcs.gla.ac.uk/mira/workshops/dublin/procs), one by S. Gabrielli and Mizzaro [5] and one by J. Reid and Mizzaro (http://www.dcs.gla.ac.uk/mira/workshops/dublin/procs/mr.pdf). This study should lead to some transformation functions (from the agreement for one kind of judgment to the agreement for another kind of judgment) for comparing the agreements obtained using different kinds of judgments.

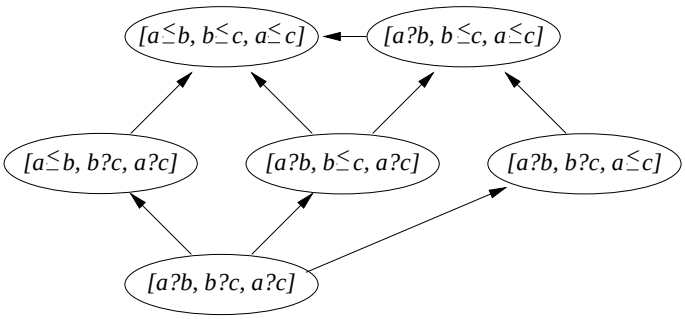

Figure 8: Some of the elements of $\Omega(\{a, b, c\})$.

Acknowledgments

I would like to thank Gianni Amati and Fabio Crestani, for making me realize that we did not agree on what relevance judgments agreement is, Andrea Fusiello for having read a draft of this paper and provided useful suggestions, Marino Miculan for inspiring discussions, and three anonymous referees for their constructive comments. This work was supported in part by a grant from the Italian National Research Council (grant nr. 210.15.11).

References

[1] G. C. Barhydt. A comparison of relevance assessments by three types of evaluator. In *Proceedings of the American Documentation Institute*, pages 383–385, Washington, DC, 1964. American Documentation Institute.

[2] G. C. Barhydt. The effectiveness of non-user relevance assessments. *Journal of Documentation*, 23(2 and 3):146–149 and 251, 1967.

[3] R. Burgin. Variations in relevance judgments and the evaluation of retrieval performance. *Information Processing & Management*, 28(5):619–627, 1992.

[4] R. C. Figueiredo. Estudo comparativo de julgamentos de relevância do usuário e não-usuário de serviços de D.S.I. *Ciéncia da Informação–Rio de Janeiro*, 7:69–78, 1978.

[5] S. Gabrielli and S. Mizzaro. Negotiating a multidimensional framework for relevance space. In this volume, 1999. Available also 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/04/99.

[6] S. P. Harter. Variations in relevance assessments and the measurement of retrieval effectiveness. *Journal of the American Society for Information Science*, 47(1):37–49, 1996.

[7] J. M. Hoffman. Experimental design for measuring the intra- and inter-group consistence of human judgment for relevance. Master's thesis, Georgia Institute of Technology, Atlanta, Georgia, 1965.

[8] J. W. Janes. Other people's judgments: A comparison of user's and other's judgments of document relevance, topicality, and utility. *Journal of the American Society for Information Science*, 45(3):160–171, April 1994.

[9] J. W. Janes and R. McKinney. Relevance judgments of actual users and secondary judges. *Library Quarterly*, 62:150–168, 1992.

[10] T. V. Kazhdan. Effects of subjective expert evaluation of relevance on the performance parameters of a document-based information-retrieval system. *Nauchno-Tekhnicheskaya Informatsiya, Seriya 2*, 13:21–24, 1979.

[11] M. E. Lesk and G. Salton. Relevance assessments and retrieval system evaluation. *Information Storage and Retrieval*, 4(3):343–359, 1968.

[12] 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', a prize for the best paper by an European student in the period 1st November 1996 - 1st November 1997.

[13] J. O'Connor. Relevance disagreements and unclear request forms. *American Documentation*, 18(3):165–177, 1967.

[14] J. O'Connor. Some independent agreements and resolved disagreements about answer-providing documents. *American Documentation*, 20(4):311–319, 1969.

[15] M. E. Rorvig. The simple scalability of documents. *Journal of the American Society for Information Science*, 41(8):590–598, 1990.

[16] Y. Y. Yao. Measuring retrieval effectiveness based on user preference of documents. *Journal of the American Society for Information Science*, 46(2):133–145, 1995.