

On the Effect of Ranking Axioms on IR Evaluation Metrics

Fernando Giner

fginer3@gmail.com

E.T.S.I. Informática UNED

Madrid, Spain

ABSTRACT

The study of IR evaluation metrics through axiomatic analysis enables a better understanding of their numerical properties. Some works have modelled the effectiveness of retrieval metrics with axioms that capture desirable properties on the set of rankings of documents. This paper formally explores the effect of these ranking axioms on the numerical values of some IR evaluation metrics. It focuses on the set of ranked lists of documents with multigrade relevance. The possible orderings in this set are derived from three commonly accepted ranking axioms on retrieval metrics; then, they are classified by their latticial properties. When relevant documents are prioritised, a subset of document rankings are identified: the *join-irreducible* elements, which have some resemblance to the concept of basis in vector space. It is possible to compute the precision, recall, *RBP* or *DCG* values of any ranking from their values in the join-irreducible elements. However this is not the case when the swapping of documents is considered.

CCS CONCEPTS

• Information systems → Retrieval effectiveness.

KEYWORDS

information retrieval, evaluation metric, lattice theory

ACM Reference Format:

Fernando Giner. 2022. On the Effect of Ranking Axioms on IR Evaluation Metrics. In *Proceedings of Proceedings of the 2022 ACM SIGIR International Conference on the Theory of Information Retrieval (ICTIR '22)*. ACM, New York, NY, USA, 11 pages. <https://doi.org/10.1145/3539813.3545153>

1 INTRODUCTION

In the Cranfield paradigm [15], relevance judgements are assigned to a ranked list of documents for a topic. Then, an evaluation metric can be seen as a mapping that relates the set of possible lists of judged documents with a numeric structure, for instance, the set of real numbers. In this way, an evaluation metric assigns numbers (scores) that correspond to or 'preserve' certain observed relations between the set of possible lists of documents.

Many properties that an evaluation metric possesses are directly related to the properties of the set of rankings where it is defined, for example, in the statement "an evaluation metric is an ordinal scale",

the evaluation metric is preserving the *existing ordering* between rankings of documents. This characteristic has been observed in some axiomatic studies, such as *linear algebra*, where a linear form, defined from a vector space to the real numbers, verifies that the image of any vector is determined by the image of a basis. It can also be seen in *measurement theory*, where the representation theorems look for the requirements in the set where a measure is defined [33, 34, 42]. In these studies the set where the mapping is defined (empirical domain), endowed with relationships and/or operations, is distinguished from the numerical values that the mapping can reach (numerical image).

The empirical domain of the IR evaluation metrics can be shaped by desirable properties, for instance, the property "it is preferred to retrieve relevant documents in the top ranking positions" establishes a preference relation in the set of rankings of documents. The use of formal properties or axioms has been successfully applied to improve and understand evaluation metrics [5, 7, 11, 25, 37, 48, 49]. The axiomatic approach of "evaluating evaluation" consists of describing desirable properties that a retrieval metric should satisfy, then the acceptability of a given metric is deduced or an evaluation metric can be axiomatically derived. We follow this axiomatic view and formally explore the effect of ranking axioms on the numerical properties of some IR evaluation metrics.

In this paper, intuitive and commonly accepted ranking axioms are considered, to ground the behaviour of IR evaluation metrics in a representative and well understood axiomatic basis. Then, these ranking axioms are operationalised and the possible orderings in the set of rankings of documents are derived. Applying results of lattice theory, a remarkable subset of elements is identified: the *join-irreducible* elements. When relevant documents are prioritised, every ranked list of documents can be expressed in a unique manner as a join-combination of join-irreducible elements, similarly to the concept of *basis* in vector space. In this context, IR evaluation metrics are real-mappings defined on the set of rankings. It is shown that precision, recall, *DCG* and *RBP* verify a kind of linearity, i.e., they are *valuations* (in lattice theory terms). This property allows to determine their values on any ranking from the values on the join-irreducible elements; similarly to how the image of a linear form is determined by the images of the elements of a basis. However, when the swapping of documents is considered, they do not necessarily verify this property. In addition to understanding how IR evaluation metrics take effect on the set of output system runs, this result has some theoretical and practical consequences, especially in studies devoted to test the behaviour of retrieval metrics.

This paper is organized as follows: In Section 2, some related work is reported, and three operations on the set of ranked lists of documents are stated. In Section 3, the basic notions of lattice theory are introduced. Section 4 formalizes the problem in the context of IR. Sections 5 and 6 contain the main results, where the possible

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.
ICTIR '22, July 11–12, 2022, Madrid, Spain.

© 2022 Copyright held by the owner/author(s). Publication rights licensed to ACM.
ACM ISBN 978-1-4503-9412-3/22/07...\$15.00
<https://doi.org/10.1145/3539813.3545153>

orderings on the set of rankings of documents and their structural properties are derived for the set-based and rank-based retrieval. Finally, in Section 7 some conclusions and applications are drawn.

2 RELATED WORK

The formal analysis of IR evaluation metrics has contributed to a better understanding of their properties. One of the first attempts is given in [51], where the effectiveness and the efficiency of IR evaluation metrics are analyzed in terms of a 2-by-2 contingency table of pertinence and retrieval. Later, van Rijsbergen [9, 52], tackle the issue of the foundations of measurement in IR through a conjoint (additive) structure based on precision and recall; then, he examines the properties of a measure on this prec-recall structure. In [12], a similar conjoint structure is defined, but on the contingency table of the binary retrieval; then, they study the properties of the proposed MZ-metric. In [11], two axioms are highlighted, then, all the IR metrics, which are compliant with them, are expressed as a linear combination of the number of relevant/non-relevant retrieved documents. In [54], user judgements on documents are formally described by a weak order, then, a measure of system performance is presented, whose appropriateness is demonstrated through an axiomatic approach. In [30], a framework for the theoretical comparison of IR models, based on situation theory, is presented. It enables an inference mechanism with the axiomatised concept of *aboutness*. In [4, 5], some properties of evaluation metrics for clustering are formalized, describing how IR metrics should behave. In [37], it is proposed a formal framework based on numerical properties of IR evaluation metrics, it helps to better understand the relative merits of IR metrics for different applications. In [48], the axioms that an evaluation measure for classification should satisfy are discussed, then the K evaluation metric is proposed. In [13, 35] an axiomatic definition of IR effectiveness metric is provided, in a unifying framework for ranking, classification and clustering. In [49], it is discussed what properties should enjoy an evaluation measure for quantification.

In recent years, some forums and conferences [1, 2, 13, 35, 55] have encouraged the search for desirable properties expressed mathematically as formal constraints to assess the optimality of retrieval models. They are formal properties that a “good” retrieval model should fulfil [20–23]. This analytical approach allows to explore retrieval models and how best to improve them, in order to achieve higher retrieval effectiveness. It has been successfully applied to the study of basic models [19, 20], pseudo-relevance feedback methods [16, 17, 40], translation retrieval models [31, 41] and neural network retrieval models [44].

In [6], a survey about the attempts to formalize the properties of evaluation metrics is provided. They consider a wide variety of information access tasks and classify these properties into four categories: (i) general axioms [37, 48], (ii) axioms for classification metrics [3, 48, 50], (iii) axioms for clustering metrics [4, 18, 36, 43] and (iv) ranking axioms [5, 7, 11, 25, 37].

2.1 Ranking Axioms

In terms of the batch evaluation of IR systems, some works have highlighted formal properties in the set of rankings. In [11], two axioms are defined: the *monotonicity* and the *Archimedian*, which

determine the form of the evaluation metrics from a measurement theoretic perspective. In [37], the evaluation metrics are characterized by seven numeric properties, such as *convergence*, which states that relevant documents must occur above non-relevant ones, *top-weightedness*, *localization* and others. It provides a framework to classify the metrics according to their effectiveness. In [5, 7], the following properties are highlighted: *priority*, which states that swapping documents in a correct way is a preferred option, *deepness*, which affirm that deeper positions in the ranking are less likely to be explored, *closeness threshold*, which states that there is an area at the top of the ranking which is always explored, *confidence*, etc. They derive evaluation metrics, which can be applied to several information access tasks. In [26], the set of rankings of documents is characterized by the *replacement* operation, which prioritize the replacement of a document by another document with higher relevance degree; and the *swap* operation, which states that swapping a less relevant document with a more relevant one in a lower rank position is a preferred option. In [6], the swapping of documents is highlighted since it is common to many other tasks.

We conduct our analysis on the basis of three formal properties, which express preferences on the empirical domain of the retrieval metrics. They are explicitly or implicitly present in most of these works: (i) *relevance*: “a run retrieving relevant documents is preferred to another one retrieving less relevant documents”, this property has been studied under several names such as *monotonicity* [11], *replacement* [25] or *confidence* [7]; (ii) *deepness*: “it is better one relevant document in the highest part of a list than many other relevant documents in the bottom part”, this property is stated in [5, 7] as a *deepness threshold* or *deepness* and they assert that one relevant document in the highest part of a list is preferred over many other relevant documents in the bottom part; and (iii) *swapping*: “swapping a less relevant document in a higher rank position with a more relevant one in a lower rank position is a preferred option”, this property is known as *convergence* [37], *swap* [6, 25] or *priority constraint* [5].

These ranking axioms can be expressed in terms of ranking operations [29, 53], it enables to explore to what extent the numerical properties of IR evaluation metrics can be explained by the effect of ranking axioms, and which axioms apply in these situations. This characteristic has been shown in [26], where the orderings between rankings are based on two operations: *replacement* and *swap*. In this paper, the ranking constraints are operationalized as Boolean predicates which, given a pair of ranked lists of documents, express a preference between them.

Operation 1 (Replacement):

“Replacing a document for another one with a higher relevance degree is a preferred option”.

This operation implements the *relevance* axiom and it establishes a simple property: *prioritizing relevant documents*.

Operation 2 (Projection):

“Removing all documents that are not considered in the highest or more relevant part of a list is a preferred option”.

This operation implements the *deepness* axiom and it remarks on a specific part (the highest or more notable) where it should be

considered a relevant document, without paying attention to the rest of the documents.

Operation 3 (Swapping):

“Swapping a less relevant document in a higher rank position with a more relevant one in a lower rank position is a preferred option”.

This operation implements the *swapping* axiom.

These three operations express preferences between ranked lists of documents, thus, every operation give rise to a total or partial ordering in the empirical domain of IR evaluation metrics. One of the major advantages of having operationalized the axioms is that it facilitates the combinations of them. In Sections 5 and 6, the possible combinations of these operations are considered, resulting in five different orderings in the set of ranked lists of documents.

3 LATTICE THEORY

In this section we recall some basic notions of lattice theory, for additional background, we refer the reader to the textbooks [8, 28].

A *partially ordered set* or *poset* is a non-empty set, X , together with a binary, reflexive, antisymmetric and transitive relation, \leq , defined on X . A poset (X, \leq) is called a *totally ordered set* or a *chain* if any two elements, $x, y \in X$, are comparable, i.e., $x \leq y$ or $y \leq x$.

A (closed) interval is a subset of X defined as $[x, y] = \{z \in X : x \leq z \leq y\}$. It is said that y *covers* x in X , denoted by $x \lessdot y$, if there does not exist $z \in X$ such that $x \neq z \neq y$ and $x \leq z \leq y$ ¹.

Thus, (X, \leq) has an associated graph, called the *Hasse diagram*, where nodes are labelled with the elements of X and the edges indicate the covering relation. Nodes are represented in different levels, if y covers x , then x is below y in the diagram.

A noteworthy property of a finite poset is that its covering relation is the transitive reduction of the partial order relation, i.e., the covering relation cannot contain the following pairs $x \lessdot y, y \lessdot z$ and $x \lessdot z$ at the same time, for any $x, y, z \in X$.

An upper bound for a subset, $S \subseteq X$, is an element, $x \in X$, for which $s \leq x, \forall s \in S$. A lower bound for a subset of a poset is defined analogously. The greatest lower bound or *meet* of two elements $x, y \in X$ is denoted by $x \wedge y$. The least upper bound or *join* of two elements $x, y \in X$ is denoted by $x \vee y$. It is said that X has a *top element*, denoted by $1 \in X$, if $x \leq 1, \forall x \in X$; dually, $0 \in X$ is the *bottom element*, if $0 \leq x, \forall x \in X$.

Considering the join and the meet as operators, (X, \wedge, \vee, \leq) is a *lattice* if (X, \leq) is a poset and every two elements of X have a meet and a join. A *sublattice*, $S \subseteq X$, is a subset of the lattice that is closed to the meet and join operations, i.e., $x \wedge y, x \vee y \in X, \forall x, y \in S$.

A lattice is *distributive* if its meet operator distributes over its join operator, i.e., $x \wedge (y \vee z) = (x \wedge y) \vee (x \wedge z), \forall x, y, z \in X$.

For instance, any chain is a distributive lattice [8]. There are two specific lattices that characterize the distributive lattices, the diamond lattice (M_3) and the pentagon lattice (N_5), which are shown in Figure 1.

¹A binary relation of order can be described in two ways. For example, the relation “two real numbers are related if one of them is, at least, one higher unity than the other”, can be described: (1) in an extensive way: $a \leq b \Leftrightarrow a \leq b + 1, \forall a, b \in \mathbb{R}$, i.e., describing all the possible pairs which are related; or (2) with the covering relation: $a < b \Leftrightarrow a = b + 1, \forall a, b \in \mathbb{R}$, i.e. describing only the pairs which one element covers the other.

THEOREM 1 (BIRKHOFF [8]). *A lattice, X , is distributive, if and only if it does not contain M_3 or N_5 as sublattices.*

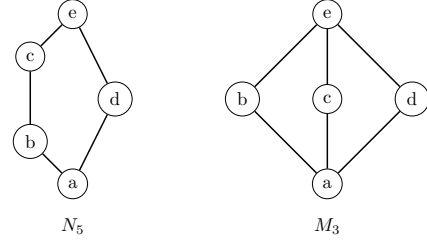


Figure 1: Lattices N_5 and M_3 are not distributive

There are some remarkable elements in a lattice, given $j \in X$ (with $j \neq 0$), it is said to be a *join-irreducible* element, if $x \vee y = j$ implies $x = j$ or $y = j$. Thus, $j \in X$ is join-irreducible if it cannot be expressed as the join of two elements that are strictly smaller than j . We denote the set of join-irreducible elements of X by J_X ². The following result highlights the importance of the join-irreducible elements of a finite distributive lattice.

THEOREM 2 (BLYTH [10]). *If X is a finite distributive lattice, then every element of $X \setminus \{0\}$ can be expressed uniquely as an irredundant join of join-irreducible elements.*

Join-irreducible elements have some resemblance to the concept of basis in vector space since every element is determined by them.

A real-valued mapping defined on a lattice, $v : X \rightarrow \mathbb{R}$, is a *valuation*, if it satisfies the following property: $v(x) + v(y) = v(x \vee y) + v(x \wedge y), \forall x, y \in X$.

The next result shows a characterization of the valuations in terms of the join-irreducible elements.

THEOREM 3 (ROTA [45]). *A valuation, defined on a finite distributive lattice X , is uniquely determined by the values it takes on the set of join-irreducible elements of X .*

4 IR SETTING

In this section, some notation needed throughout the paper is introduced. The formalism of [26] is adopted.

Consider a finite set of documents that are retrieved from query terms provided by a user. An (ordered or not) collection of N retrieved documents is called the *system run* of length N for a topic.

Once documents have been retrieved, they can be classified by their relevance degree. Let (REL, \leq) be a finite and totally ordered set of relevance degrees, where $REL = \{a_0, \dots, a_c\}$ with $a_i < a_{i+1}$, for each $i \in \{1, \dots, c-1\}$. These relevance degrees can be categorical labels. To handle numerical values, a *gain function*, $g : REL \rightarrow \mathbb{R}^+$, is considered as a map that assigns a positive real number to any relevance degree. In this paper, it can be assumed that $g(a_0) = 0$ and g is a strictly increasing function.

When every retrieved document of a system run is classified with a relevance degree, a *judged run* is obtained, denoted by \hat{r} .

²The join-irreducible elements are those that have only one descendant in the covering graph [28]

In *set-based retrieval*, each judged run is an unordered set of N relevance degrees, which may contain the same element several times. For instance, for $N = 4$ and 3 relevance degrees, a judged run is given by $\hat{r} = \{a_2, a_1, a_1, a_0\}$. For the sake of clarity, the convention is used to represent \hat{r} as $\{\hat{r}_1, \dots, \hat{r}_N\}$, where $\hat{r}_i \geq \hat{r}_{i+1}$, for any $i \in \{1, \dots, N-1\}$, i.e., the relevance degrees are listed in decreasing order. This process does not affect the results obtained. The j -th element of the set \hat{r} is denoted by \hat{r}_j .

In *rank-based retrieval*, each judged run is an ordered list of N relevance degrees. For instance, for $N = 4$ and 3 relevance degrees, a judged run is given by $\hat{r} = (a_1, a_0, a_2, a_1)$. In this case, the j -th element of \hat{r} is denoted by $\hat{r}[j]$.

The set of all the possible judged runs of length N is a finite set, denoted by $R(N)$.

The *replacement* operation of Section 2 establishes that relevant documents are prioritized. This property can be quantified with the following concept.

DEFINITION 1. *The cumulated mass of relevance of a judged run, $\hat{r} \in R(N)$, for a relevance degree, $a_j \in REL$, is the number of documents of \hat{r} with a higher relevance degree than a_j .*

In set-based retrieval, considering that $\hat{r} = \{\hat{r}_1, \dots, \hat{r}_N\} \in R(N)$, it is given by $|\{i : \hat{r}_i \geq a_j\}|$. In rank-based retrieval, considering that $\hat{r} = (\hat{r}[1], \dots, \hat{r}[N]) \in R(N)$, it is given by $|\{i : \hat{r}[i] \geq a_j\}|$

REMARK 1. *In rank-based retrieval, this definition can be generalized to a specific position of the ranking. The cumulated mass of relevance, for $a_j \in REL$, at position k , where $1 \leq k \leq N$, is the number of documents with a higher relevance degree than a_j , from the first to the k -th position, which is given by $|\{1 \leq i \leq k : \hat{r}[i] \geq a_j\}|$*

The *projection* operation establishes that the documents that are not considered in the highest or most relevant part of the list should be removed. In rank-based retrieval, the highest or most relevant part of the list is specified by the top positions of the ranking. Thus, this property can be understood as removing those documents below a specific position in the ranking. On the other hand, in the set-based retrieval, there are no positions of a ranking, but instead the more relevant part of the list is specified by the relevance degrees. Thus, this operation can be understood as removing those documents that do not have high relevance degrees.

Considering different combinations of the three properties of Section 2, preference relations will be obtained on the set of judged runs, i.e., $R(N)$ will be endowed with an ordering, which will be denoted by \leq .

If this ordering is a poset, then $(R(N), \wedge, \vee, \leq)$ can be considered as a lattice, since $R(N)$ is a finite set, where every pair of elements have a meet and join, trivially.

An evaluation metric can be seen as a mapping between $R(N)$ and the set of real numbers. In set based retrieval some remarkable evaluation measures are as follows.

- Generalized precision (gP) [26] :

$$gP(\hat{r}) = \frac{1}{N} \cdot \sum_{i=1}^N \frac{g(\hat{r}_i)}{g(a_c)}.$$

- Generalized recall (gR) [26]

$$gR(\hat{r}) = \frac{1}{RB} \cdot \sum_{i=1}^N \frac{g(\hat{r}_i)}{g(a_c)},$$

where RB is the recall base (total number of relevant documents).

These evaluation measures can also be considered in rank-based retrieval, in addition to the following.

- Graded rank-biased precision ($gRBP$) [26, 39, 47]:

$$gRBP(\hat{r}) = \frac{(1-p)}{g(a_c)} \cdot \sum_{i=1}^N p^{i-1} \cdot g(\hat{r}[i]).$$

- Discounted cumulated gain (DCG) [32]:

$$DCG_b(\hat{r}) = \sum_{i=1}^N \frac{g(\hat{r}[i])}{\max\{1, \log_b i\}},$$

In the two following sections, the possible orderings in set-based and rank-based retrieval are deduced by means of possible combinations of the three operations presented (see Section 2). It has to be noted that considering multiple operations at the same time, could yield a set of possibly conflicting preferences for ranked lists of documents. In this case, those combinations of operations that generate an ordering relation with cycles are dismissed since they do not satisfy the transitivity of a partial order. Likewise, those combinations of axioms that generate a disconnected Hasse diagram, i.e., the cases where there exists an element or a group of elements that cannot be compared with the rest of the elements, are also excluded. In these cases, the obtained results can be applied to each connected component of the Hasse diagram.

5 SET-BASED RETRIEVAL

In this scenario, a system run is a set of relevance degrees, where the judged documents are listed in decreasing order.

It is not possible to consider the *swapping* operation since interchanging documents in a set has no effect. On the other hand, the *replacement* operation prioritizes relevant documents, and the *projection* operation focuses on the most relevant part of the list.

Thus, the possible combinations of operations will be: (i) considering the *projection* and the *replacement* operations jointly and (ii) considering the *replacement* operation individually.

Note that it is not possible to consider the *projection* operation individually since it refers to a specific part of the list, and the *replacement* operation is needed to prioritize relevant documents.

5.1 Projection+Replacement [set-based]

Given a pair of judged runs, the *projection* operation selects the more relevant part of the lists, i.e., the highest relevance degree at which the two system runs differ. In addition, considering the *replacement* operation, the ranking that has greater *cumulative mass of relevance* (or equivalently more relevant documents³) for that relevance degree is preferred. Formally:

³Considering the highest relevance degree at which the two system runs differ, the difference between the *cumulated mass of relevance* is determined by the number of documents with that relevance degree.

Projection+Replacement [set-based]

Let $\hat{r}, \hat{s} \in R(N)$ such that $\hat{r} \neq \hat{s}$, and let k be the largest relevance degree at which the two runs differ for the first time, i.e., $k = \max \{j \leq c : |\{i : \hat{r}_i = a_j\}| \neq |\{i : \hat{s}_i = a_j\}|\}$, then they are ordered by:

$$\hat{r} \leq \hat{s} \iff |\{i : \hat{r}_i = a_k\}| \leq |\{i : \hat{s}_i = a_k\}|$$

A system run is preferred to another, if for the highest relevance degree at which the two systems runs differ, the first has more documents with this relevance degree.

$(R(N), \leq)$ is a totally ordered set or a chain⁴. Thus, the meet and join operations of a pair of elements are the minimum and maximum of both, respectively, i.e., $\hat{r} \wedge \hat{s} = \min\{\hat{r}, \hat{s}\}$, $\hat{r} \vee \hat{s} = \max\{\hat{r}, \hat{s}\}$.

As noted in Section 3, $R(N)$ endowed with these operations is a distributive lattice since it is a chain. Moreover, every element of $R(N)$ is join-irreducible, except $\hat{0}$, since these elements have only one descendant in the Hasse diagram, $J_{R(N)} = R(N) \setminus \{\hat{0}\}$. Then, it is possible to apply Theorem 3, which states that the values reached by an evaluation metric are determined by the values of the join-irreducible elements. However, this is a trivial case where no relevant information is obtained since the values reached by an evaluation measure are determined by every ranking of the empirical domain, except $\hat{0}$.

5.2 Replacement [set-based]

Considering the *replacement* operation individually, the relevant documents have to be prioritized, i.e., a system run is preferred to another, if it has more *cumulated mass of relevance* for any relevance degree. Formally:

Replacement [set-based]

Any pair of system runs, $\hat{r}, \hat{s} \in R(N)$, such that $\hat{r} \neq \hat{s}$, is ordered by:

$$\hat{r} \leq \hat{s} \iff |\{i : \hat{r}_i \geq a_j\}| \leq |\{i : \hat{s}_i \geq a_j\}|, \forall j \in \{0, \dots, c\}$$

This ordering considers a run greater than another if, fixing any relevance degree, it has more documents above that relevance degree than does the other.

In this ordering not every pair of rankings is comparable, for instance, $\hat{r} = \{a_2, a_2, a_0\}$ and $\hat{s} = \{a_3, a_0, a_0\}$, since $|\{i : \hat{r}_i \geq a_3\}| = 0 < 1 = |\{i : \hat{s}_i \geq a_3\}|$, while $|\{i : \hat{r}_i \geq a_2\}| = 2 > 1 = |\{i : \hat{s}_i \geq a_2\}|$.

However, $R(N)$ endowed with this ordering is a *poset*⁵. Figure 2 shows the Hasse diagram for the case $c = 2$ and $N = 2$.

Considering the convention to represent judged runs in decreasing order, the join and the meet operations have an explicit expression.

PROPOSITION 1. Given $\hat{r} = \{\hat{r}_1, \dots, \hat{r}_N\}$ and $\hat{s} = \{\hat{s}_1, \dots, \hat{s}_N\}$ in $(R(N), \wedge, \vee, \leq)$, with the **replacement** [set-based] partial order, their

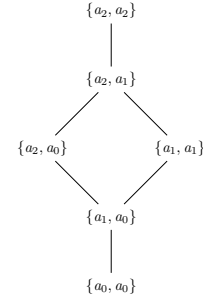


Figure 2: Hasse diagram for $c = 2$ and $N = 2$ in the set-based retrieval

meet and join can be expressed as:

$$\begin{aligned} \hat{r} \wedge \hat{s} &= \left\{ \min\{\hat{r}_1, \hat{s}_1\}, \dots, \min\{\hat{r}_N, \hat{s}_N\} \right\} \\ \hat{r} \vee \hat{s} &= \left\{ \max\{\hat{r}_1, \hat{s}_1\}, \dots, \max\{\hat{r}_N, \hat{s}_N\} \right\} \end{aligned}$$

In this case, $(R(N), \leq)$, endowed with the meet and join operations, is trivially a lattice, since $R(N)$ is a finite set. The following result shows the structure of $R(N)$.

PROPOSITION 2. $(R(N), \wedge, \vee, \leq)$ is a distributive lattice, where \leq is the **replacement** [set-based] partial order.

In this context, a remarkable subset of judged runs is identified by the following result.

PROPOSITION 3. The join-irreducible elements of $(R(N), \wedge, \vee, \leq)$, where \leq is the **replacement** [set-based] ordering, are:

$$\begin{array}{cccc} \{a_1, a_0, \dots, a_0\}, & \{a_1, a_1, a_0, \dots, a_0\}, & \dots, & \{a_1, \dots, a_1\}, \\ \{a_2, a_0, \dots, a_0\}, & \{a_2, a_2, a_0, \dots, a_0\}, & \dots, & \{a_2, \dots, a_2\}, \\ \vdots & \vdots & \vdots & \vdots \\ \{a_c, a_0, \dots, a_0\}, & \{a_c, a_c, a_0, \dots, a_0\}, & \dots, & \{a_c, \dots, a_c\} \end{array}$$

For example, for the case $N = 5$ and three relevance degrees ($c = 2$), the Hasse diagram is shown in Figure 3, where the join-irreducible elements have been marked.

Proposition 1 provides an analytical way to determine the meet and the join of two judged runs. A graphical way to obtain the same results is illustrated in Figure 3, where the join of two elements is the intersection of the minimum upwards paths starting from these elements, and the meet of two judged runs is the intersection of the minimum downwards paths.

The result obtained in Proposition 2 allows us to apply Theorem 2 to this partial order, showing the importance of the join-irreducible elements. It states that every judged run can be expressed as a join of the join-irreducible elements. This result can be checked for some judged runs of this example:

$$\begin{aligned} \hat{r}^5 &= \hat{r}^2 \vee \hat{r}^3, & \hat{r}^8 &= \hat{r}^3 \vee \hat{r}^4, & \hat{r}^{10} &= \hat{r}^4 \vee \hat{r}^7, & \hat{r}^{11} &= \hat{r}^3 \vee \hat{r}^6, \\ \hat{r}^{14} &= \hat{r}^3 \vee \hat{r}^9, & \hat{r}^{15} &= \hat{r}^6 \vee \hat{r}^{12}, & \hat{r}^{16} &= \hat{r}^7 \vee \hat{r}^9, & \hat{r}^{18} &= \hat{r}^9 \vee \hat{r}^{12} \end{aligned}$$

Once the properties of the empirical domain have been analysed, it is possible to study the properties of the evaluation measures defined on it. Figure 4 shows the numerical values reached by the generalized precision and recall (they are the same) for this example.

⁴This ordering is the same as the “Total Ordering” (set-based case) studied in [26], where it can be seen the linearity of the order.

⁵This partial order is the same as the “Partial Ordering” (set-based case) studied in [26], where it can be seen this result.

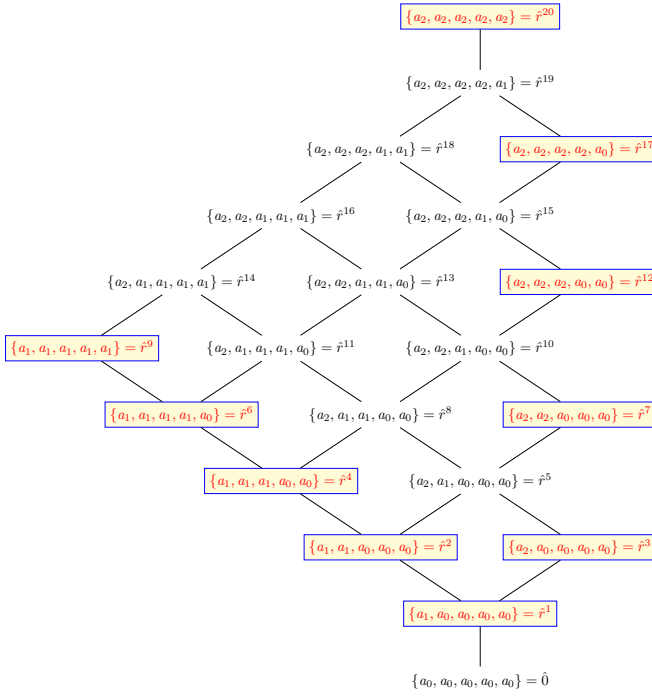


Figure 3: Hasse diagram for $c = 2$ and $N = 5$ in the set-based retrieval

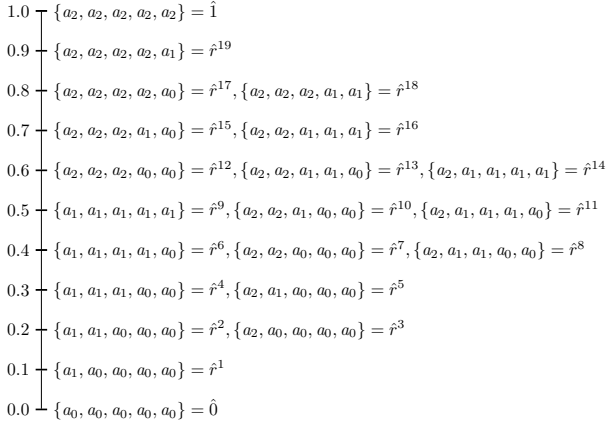


Figure 4: Generalized Precision and Recall for $c = 2$ and $N = 5$ in the set-based retrieval

These values have been obtained with the definition of gP in Section 4.

In Section 3, it has been shown that if a real mapping verifies a kind of linearity, i.e., if it is a valuation, then it is completely determined by the values of the join-irreducible elements. The following result shows that the IR evaluation metrics of this scenario verify this property.

PROPOSITION 4. *The generalized precision and recall are valuations, when they are defined on $(R(N), \wedge, \vee, \leq)$, where \leq is the relevance [set-based] partial order.*

The main result of this subsection is obtained from Propositions 2 and 4, which allow to apply Theorem 3 to this partial order. Thus, the values reached by gP and gR are determined by the values reached on the set of join-irreducible elements.

Let us now look at some practical aspects of calculating these values. By means of Proposition 4, the numerical values of gP and gR can be computed with the characterization property of valuations: $v(\hat{r} \vee \hat{s}) = v(\hat{r}) + v(\hat{s}) - v(\hat{r} \wedge \hat{s})$. Note that the last operand of the second part of the equality contains a meet of two join-irreducible elements. Considering Propositions 1 and 3, it can be easily seen that every meet of any pair of join-irreducible elements is a join-irreducible. Thus, the characterization of valuations is sufficient to determine any numerical value.

For example, Figure 4 shows that $gP(\hat{r}^5) = 0.3$. This value can also be obtained from the values of gP in the set of join-irreducible elements. Making use of Theorem 2 and Proposition 4, $gP(\hat{r}^5) = gP(\hat{r}^2 \vee \hat{r}^3) = gP(\hat{r}^2) + gP(\hat{r}^3) - gP(\hat{r}^2 \wedge \hat{r}^3) = 0.2 + 0.2 - gP(\hat{r}^1) = 0.2 + 0.2 - 0.1 = 0.3$.

The following equalities check some numerical values of the gP evaluation measure.

$$\begin{aligned} gP(\hat{r}^5) &= 0.2 + 0.2 - 0.1 = 0.3 & gP(\hat{r}^8) &= 0.2 + 0.3 - 0.1 = 0.4 \\ gP(\hat{r}^{10}) &= 0.3 + 0.4 - 0.2 = 0.5 & gP(\hat{r}^{11}) &= 0.2 + 0.4 - 0.1 = 0.5 \\ gP(\hat{r}^{13}) &= 0.4 + 0.4 - 0.2 = 0.6 & gP(\hat{r}^{14}) &= 0.2 + 0.5 - 0.1 = 0.6 \end{aligned}$$

Therefore, the values of gP and gR are determined by their values on the set of join-irreducible elements.

6 RANK-BASED RETRIEVAL

In this scenario, a judged run is an array of relevance degrees. Considering the *swapping* operation individually will lead to a disconnected Hasse diagram. For example, the same result is always obtained when it is applied to (a_1, \dots, a_1) ; then, this element will not be comparable to any other judged run of $R(N)$. The same can be said for the *projection* operation, which selects only specific positions in the ranking. On the other hand, the *replacement* operation should be present in any combination since prioritizing relevant documents is a requirement of any retrieval system.

Then, it is possible to consider the following combinations: (i) the *projection* and *replacement* operations jointly and (ii) the *replacement* operation individually.

To combine the *swapping* operation with other operations, it must be noted that the covering relation cannot contain the following pairs $\hat{r} \prec \hat{s}$, $\hat{s} \prec \hat{t}$ and $\hat{r} \prec \hat{t}$ at the same time (since it is a transitive reduction; see Section 3). In fact, when the *swapping* operation between any pair of positions is combined with the *replacement* operation in any position of the ranking, it leads to binary relations that do not satisfy this condition. For example, for $R(3)$ in the binary case, $\hat{r} = (a_1, a_0, a_0)$ covers $\hat{0} = (a_0, a_0, a_0)$ when considering *replacement* in the first position. However, $\hat{0} \prec (a_0, a_1, a_0) \prec \hat{r}$ when considering *swapping* in the first two positions.

This means that to combine the *swapping* and the *replacement* operations, some restriction should be imposed on the *replacement*

property, for example, prioritizing the relevant documents only in some ranking positions. In this case, note that *replacement* should be considered only in the last positions of the ranking since otherwise situations in which the covering relation is not a transitive reduction of its partial order could occur; the example in the previous paragraph illustrates this situation.

Thus, the combination of (iii) the *swapping* with the *replacement* operation only at the last position is also considered.

Finally, combining the three operations, i.e., *swapping*, *replacement* only at the last position and *projection*, would lead to a poset whose covering relation is not a transitive reduction of its partial order. For example, for $R(3)$ in the binary case, $\hat{s} = (a_1, a_0, a_0)$ covers $\hat{r} = (a_0, a_1, a_0)$ when considering *swapping* between the first and second positions. However, considering *replacement* in the third position, $\hat{r} \prec (a_0, a_1, a_1) = \hat{t}$, and when applying the *projection* operation $\hat{t} \prec \hat{s}$ would be obtained.

6.1 Projection+Replacement [rank-based]

Given a pair of judged runs, the *projection* operation selects the more relevant part of the ranking, i.e., the highest-ranking position at which the two system runs differ. In addition, if the *replacement* operation is also considered, the ranking with the highest *cumulative mass of relevance* (or equivalently, the highest relevance degree⁶) at that position is preferred. Formally:

Projection+Replacement [rank-based]

Let $\hat{r}, \hat{s} \in R(N)$ such that $\hat{r} \neq \hat{s}$; then, there exists $k = \min\{j \leq N : \hat{r}[j] \neq \hat{s}[j]\}$. Any pair of distinct system runs is ordered by:

$$\hat{r} \leq \hat{s} \iff \hat{r}[k] \leq \hat{s}[k]$$

A system run is preferred to another if it has a higher relevance degree in the highest-rank position at which the two system runs differ.

$(R(N), \leq)$ is a totally ordered set or a chain⁷. Thus, this is a trivial case as in the **projection+replacement** [set-based] total order.

6.2 Replacement [rank-based]

Considering the *replacement* operation individually, the relevant documents must be prioritized, i.e., a system run is preferred to another if it has more *cumulated mass of relevance*, for any relevance degree. Formally:

Replacement [rank-based]

Any pair of system runs, $\hat{r}, \hat{s} \in R(N)$, such that $\hat{r} \neq \hat{s}$, is ordered by:

$$\hat{r} \leq \hat{s} \iff \left| \{i : \hat{r}[i] \geq a_j\} \right| \leq \left| \{i : \hat{s}[i] \geq a_j\} \right|, \forall j \in \{0, \dots, c\}$$

REMARK 2. *Considering the cumulative mass of relevance, this ordering can be equivalently described with the following binary relation. Let $\hat{r}, \hat{s} \in R(N)$ be two judged runs:*

$$\hat{r} \leq \hat{s} \iff \hat{r}[k] \leq \hat{s}[k], \quad \forall k \in \{1, \dots, N\}$$

⁶Considering the highest ranking position at which the two system runs differ, the difference between the *cumulated mass of relevance* is determined by the relevance degree at that position.

⁷This partial order is the same as the “Total Ordering” (rank-based case) studied in [26], where it can be seen the linearity of the ordering.

A judged run is greater than another if, at each rank position, it has higher relevance degrees.

In this ordering not every judged run is comparable. For instance, for $N = 3$ in the binary case, $\hat{r} = (a_1, a_0, a_0)$ and $\hat{s} = (a_0, a_1, a_0)$ are not comparable, since $\hat{s}[1] \leq \hat{r}[1]$, but $\hat{r}[2] \leq \hat{s}[2]$.

However, this ordering is a poset since (REL, \leq) is a totally ordered set. Given $\hat{r}, \hat{s}, \hat{t} \in R(N)$ and considering each position in the ranking, it is reflexive since $\hat{r}[k] \leq \hat{r}[k]$; it is antisymmetric since if $\hat{r}[k] \leq \hat{s}[k]$ and $\hat{s}[k] \leq \hat{r}[k]$, then $\hat{r}[k] = \hat{s}[k]$; and it is transitive since if $\hat{r}[k] \leq \hat{s}[k]$ and $\hat{s}[k] \leq \hat{t}[k]$, then $\hat{r}[k] \leq \hat{t}[k]$.

The join and meet operations have an explicit expression.

PROPOSITION 5. *Given $\hat{r} = (\hat{r}_1, \dots, \hat{r}_N)$ and $\hat{s} = (\hat{s}_1, \dots, \hat{s}_N)$ in $(R(N), \wedge, \vee, \leq)$, with the **replacement** [rank-based] partial order, the meet and join can be expressed as:*

$$\begin{aligned} \hat{r} \wedge \hat{s} &= \left(\min\{\hat{r}_1, \hat{s}_1\}, \dots, \min\{\hat{r}_N, \hat{s}_N\} \right) \\ \hat{r} \vee \hat{s} &= \left(\max\{\hat{r}_1, \hat{s}_1\}, \dots, \max\{\hat{r}_N, \hat{s}_N\} \right) \end{aligned}$$

In this case, $(R(N), \leq)$ endowed with the meet and join operation is, trivially, a lattice since $R(N)$ is a finite set. It can be said more about the structure of $R(N)$ as it is shown in the following result.

PROPOSITION 6. *$(R(N), \wedge, \vee, \leq)$, with the **replacement** [rank-based] partial order, is a distributive lattice.*

In this context, a remarkable subset of judged runs is identified with the following result.

PROPOSITION 7. *The join-irreducible elements of $(R(N), \wedge, \vee, \leq)$, where \leq is the **replacement** [rank-based] ordering, are:*

$$\begin{array}{cccc} (a_1, a_0, \dots, a_0), & (a_0, a_1, a_0, \dots, a_0), & \dots, & (a_0, a_0, \dots, a_1), \\ (a_2, a_0, \dots, a_0), & (a_0, a_2, a_0, \dots, a_0), & \dots, & (a_0, a_0, \dots, a_2), \\ \vdots & \vdots & \vdots & \vdots \\ (a_c, a_0, \dots, a_0), & (a_0, a_c, a_0, \dots, a_0), & \dots, & (a_0, a_0, \dots, a_c) \end{array}$$

For example, for the case $N = 3$ and three relevance degrees ($c = 2$), the Hasse diagram has been split into Figures 5 and 6. It can be useful to visualize each figure in the three-dimensional space; thus, there are four cubes. Figure 6 should be placed on top of Figure 5, it can be seen that the top elements of Figure 5 coincide with the bottom elements of Figure 6. The join-irreducible elements have been marked in these figures.

The meet and join of two system runs can be graphically obtained in a similar manner as in the **replacement** [set-based] partial order. The structural properties of this partial order are obtained by applying Theorem 2; thus, every judged run can be expressed as a join of the join-irreducible elements.

Once the properties of the empirical domain have been analysed, it is possible to study the properties of the evaluation metrics defined on it. Figure 7 shows the numerical values reached by the generalized precision and recall (they are the same) for this example. These values have been obtained with the definition of gP in Section 4.

In this scenario, a noteworthy property of the generalized precision and recall, $gRBP$ and DCG_b is the following result.

PROPOSITION 8. *The generalized precision and recall, $gRBP$ and DCG_b are valuations, when they are defined on $(R(N), \wedge, \vee, \leq)$, where \leq is the **replacement** [rank-based] partial order.*

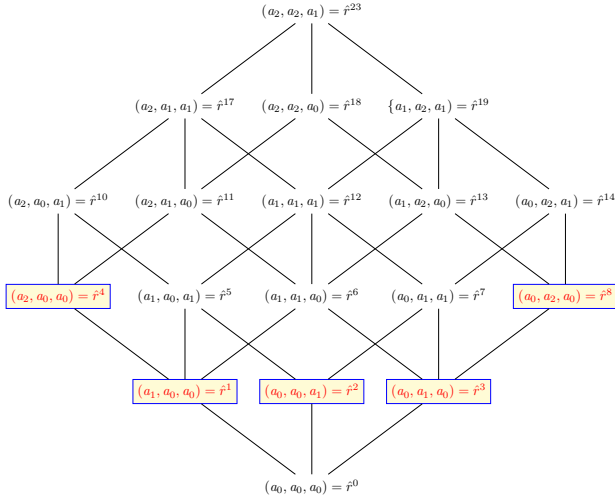


Figure 5: Hasse diagram for $c = 2$ and $N = 3$ in the rank-based retrieval, part 1

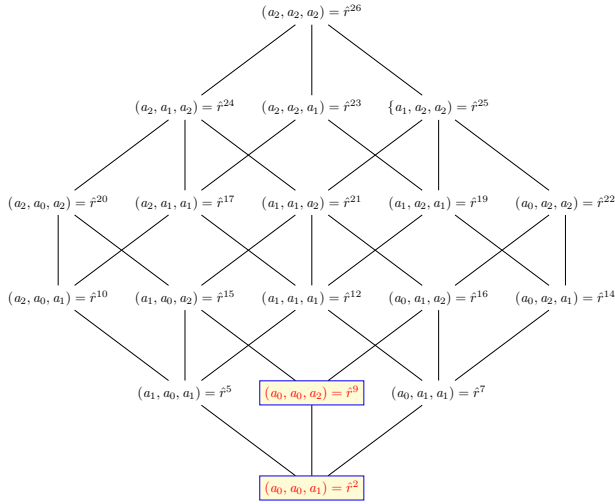


Figure 6: Hasse diagram for $c = 2$ and $N = 3$ in the rank-based retrieval, part 2

The main result of this subsection is obtained from Propositions 6 and 8, which allow to apply Theorem 3 to this partial order. Thus, the values reached by gP , gR , $gRBP$ and DCG_b are determined by the values reached on the set of join-irreducible elements.

Let us now look at some practical aspects of calculating these values. By means of Proposition 4, the numerical values of these evaluation metrics can be computed with the characterization property of valuations: $v(\hat{r} \vee \hat{s}) = v(\hat{r}) + v(\hat{s}) - v(\hat{r} \wedge \hat{s})$. Note that the last operand of the second part of the equality contains a meet of two join-irreducible elements. Considering Propositions 5 and 7, it

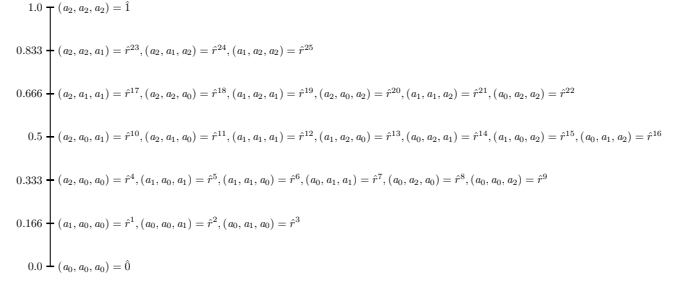


Figure 7: Generalized Precision and Recall for $c = 2$ and $N = 3$ in the rank-based retrieval

can easily be seen that the meet of any pair of join-irreducible elements is join-irreducible. Thus, the characterization of valuations is sufficient to determine any numerical value.

For example, in Figure 7, it is shown that $gP(\hat{r}^5) = 0.333$. This value can also be obtained from the values of gP in the set of join-irreducible elements. Making use of Theorem 2 and Proposition 8, $gP(\hat{r}^5) = gP(\hat{r}^1 \vee \hat{r}^2) = gP(\hat{r}^1) + gP(\hat{r}^2) - gP(\hat{r}^1 \wedge \hat{r}^2) = 0.166 + 0.166 - gP(\hat{0}) = 0.333 - 0.0 = 0.333$.

Another example where a system run is expressed with three join-irreducible elements is the following

$$gP(\hat{r}^{17}) = gP(\hat{r}^4 \vee \hat{r}^2 \vee \hat{r}^3) = gP(\hat{r}^4) + gP(\hat{r}^2 \vee \hat{r}^3) - gP(\hat{r}^4 \wedge (\hat{r}^2 \vee \hat{r}^3)).$$

As the Hasse diagram for the **replacement** [rank-based] ordering is a distributive lattice, the argument of the last operand, in the previous expression is:

$$\hat{r}^4 \wedge (\hat{r}^2 \vee \hat{r}^3) = (\hat{r}^4 \wedge \hat{r}^2) \vee (\hat{r}^4 \wedge \hat{r}^3) = \hat{0} \vee \hat{0} = \hat{0}.$$

Considering that $gP(\hat{0}) = 0$, Theorem 2 and Proposition 8:

$$gP(\hat{r}^{17}) = 0.333 + gP(\hat{r}^2 \vee \hat{r}^3) - 0 = 0.333 + gP(\hat{r}^2) + gP(\hat{r}^3) - gP(\hat{r}^2 \wedge \hat{r}^3) = 0.333 + 0.166 + 0.166 - 0 = 0.666.$$

The $gRBP$ and DCG_b evaluation metrics benefit from the same properties, we omit the checks for space limitations.

6.3 Replacement+Swapping [rank-based]

Considering the *replacement* operation, the relevant documents must be prioritized, i.e., a system run is preferred to another, if it has more *cumulated mass of relevance* for any relevance degree. In addition, when the *swapping* operation is considered, this prioritization is required at any position of the ranking since it is possible to interchange pairs of documents.

Thus, a system run is preferred to another if, for any relevance degree, it has more *cumulated mass of relevance* at any position of the ranking than the other. Formally:

Replacement+Swapping [rank-based]

Given two system runs $\hat{r}, \hat{s} \in R(N)$,

$$\hat{r} \leq \hat{s} \iff \left| \{i \leq k : \hat{r}[i] \geq a_j\} \right| \leq \left| \{i \leq k : \hat{s}[i] \geq a_j\} \right| \\ \forall j \in \{0, \dots, c\}, \forall k \in \{1, \dots, N\}$$

A run is considered larger than another one when, at each rank position, it has more relevant documents than the other up to that rank for every relevance degree.

$(R(N), \leq)$ is a partial order⁸, where not every pair of runs is comparable. For example, for $c = 2$, $N = 3$, (a_2, a_2, a_1) , and (a_2, a_2, a_0) are not comparable.

Moreover, considering $\hat{r} = (a_2, a_1, a_1)$, and $\hat{s} = (a_2, a_2, a_1)$, $\hat{r} \leq \hat{s}$ and the interval $[\hat{r}, \hat{s}]$ is a sublattice of $R(N)$, whose Hasse diagram is an N_5 [26]. Thus, by Theorem 1, $(R(N), \wedge, \vee, \leq)$ is not a distributive lattice. Consequently, not every element can be expressed in a unique manner as a join of join-irreducible elements. Thus, the values of the evaluation metrics, defined on this partial order, do not necessarily are determined by the values on the join-irreducible elements.

7 CONCLUSIONS AND APPLICABILITY

Building on some orderings on the set of ranked lists of documents, it has been verified that the properties of some IR evaluation metrics, namely, precision, recall, *DCG* and *RBP* do not depend only on its analytical expression. The empirical domain where they are defined play an important role. When the relevant documents are prioritised, these evaluation metrics are completely determined from their values on the subset of join-irreducible elements. However, this is not the case, when the swapping of documents is considered.

Our contribution is three fold: (i) the proposal grounds the behaviour of IR evaluation metrics in a representative and well understood basis of three retrieval axioms; (ii) the structural properties of the orderings are determined, in the context of lattice theory, identifying the join-irreducible elements, which have some resemblance to the concept of basis in vector space; and (iii) some numerical properties of retrieval metrics have been explained in terms of IR axioms. They can be generalized to some Carterette's models [14] and some Moffat's static user models [38], in which the conditional continuation probabilities are a function of the rank position alone.

The results of this paper have some applicability to the study of the behaviour of retrieval metrics; for instance, it has been shown that, in the **replacement** [rank-based] ordering, the total number of possible system runs of length N with c relevance degrees is c^N and the number of join-irreducible elements is $c \cdot N$. At practical level, in a rank-based retrieval of 100 documents and 3 relevance degrees, the values of precision, recall, *RBP* or *DCG* on any of the 3^{100} possible system output runs are determined by only 300 values (see Theorem 3). This result is also applicable to any other IR evaluation metric verifying the valuation condition; thus, it allows to define evaluation metrics without knowing their analytical expression. It is sufficient to assign 300 values to the join-irreducible elements. In this way, it is possible to define fine-tuned IR evaluation metrics that prioritise one type of system runs over others, by assigning a higher value to the join-irreducible elements involved in those system runs.

Another practical application is the understanding of the metric performance and the determination of bounds. Suppose that, in the rank-based retrieval of Figures 5 and 6, a system run can retrieve one document of relevance a_2 in the first position (i.e., \hat{r}^4) or 2 documents of relevance a_1 in the first and second positions (i.e., \hat{r}^6). Which of these two system runs should have a higher value for an IR evaluation metric? It seems that the system run with one document

of relevance a_2 in the second position (i.e., \hat{r}^8) should have a lower value than \hat{r}^4 ; and according to Figure 5, \hat{r}^6 should have some value in between. *RBP* verifies this property, showing its granularity; however, precision, recall and *DCG* assign the same value to the three judged runs. On the other hand, if the best achievement of a system is to retrieve one document of relevance a_2 or two documents of relevance a_1 , then the value of an IR evaluation metric on the join of \hat{r}^4 and \hat{r}^6 is a bound of system performance.

At theoretical level, determining the structural properties of the total/partial orderings derived from ranking axioms have been one of the key points of this paper. It enables to apply many results of lattice theory to the set of rankings. Considering the close relation between lattices and metric spaces [8], it is possible to classify the IR evaluation measures into metric and/or pseudo-metric mappings. In addition, taking advantage of the concept of isotone mapping and interval, it can be determined the possible ordinal and interval scales defined on these partial orders. It may open up new ways to understand ongoing debates about questions like “can we use MRR?” [24, 27, 46].

This paper has described a procedure to convert ranking axioms or operations into total/partial orders. They can be understood as ideal scenarios where some desirable properties are considered. However, these orderings cannot be considered the only ones that can be defined in the set of rankings. Considering different ranking axioms or operations could lead to other ideal scenarios which will be interesting objects to be analysed.

ACKNOWLEDGMENTS

The author thanks the anonymous reviewers for their valuable comments and suggestions.

REFERENCES

- [1] Enrique Amigó, Hui Fang, Stefano Mizzaro, and ChengXiang Zhai. 2017. Axiomatic thinking for information retrieval: And related tasks. In *Proceedings of the 40th international ACM SIGIR conference on research and development in information retrieval*. 1419–1420.
- [2] Enrique Amigó, Hui Fang, Stefano Mizzaro, and Chengxiang Zhai. 2020. Axiomatic thinking for information retrieval: introduction to special issue. *Information Retrieval Journal* 23, 3 (2020), 187–190.
- [3] Enrique Amigó, Fernando Giner, Julio Gonzalo, and Felisa Verdejo. 2017. An axiomatic account of similarity. In *Proceedings of the SIGIR'17 Workshop on Axiomatic Thinking for Information Retrieval and Related Tasks (ATIR)*.
- [4] Enrique Amigó, Julio Gonzalo, Javier Artilles, and Felisa Verdejo. 2009. A comparison of extrinsic clustering evaluation metrics based on formal constraints. *Information retrieval* 12, 4 (2009), 461–486.
- [5] Enrique Amigó, Julio Gonzalo, and Felisa Verdejo. 2013. A general evaluation measure for document organization tasks. In *Proceedings of the 36th international ACM SIGIR conference on Research and development in information retrieval*. 643–652.
- [6] Enrique Amigó and Stefano Mizzaro. 2020. On the nature of information access evaluation metrics: a unifying framework. *Information Retrieval Journal* 23, 3 (2020), 318–386.
- [7] Enrique Amigó, Damiano Spina, and Jorge Carrillo-de Albornoz. 2018. An axiomatic analysis of diversity evaluation metrics: Introducing the rank-biased utility metric. In *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*. 625–634.
- [8] Garrett Birkhoff. 1940. *Lattice theory*. Vol. 25. American Mathematical Soc.
- [9] David C. Blair. 1979. *Information Retrieval*, 2nd ed. C.J. Van Rijsbergen. London: Butterworths; 1979: 208 pp. Price: \$32.50. *JASIS* 30, 6 (1979), 374–375. <https://doi.org/10.1002/asi.4630300621>
- [10] TS Blyth. 2005. Distributive lattices. *Lattices and Ordered Algebraic Structures* (2005), 65–76.
- [11] Peter Bollmann. 1984. Two axioms for evaluation measures in information retrieval. In *SIGIR*, Vol. 84. Citeseer, 233–245.

⁸This partial order is the same as the “Partial Ordering” (rank-based case) studied in [26], where it can be seen this result.

- [12] Peter Bollmann and Vladimir S. Cherniavsky. 1980. Measurement-theoretical investigation of the MZ-metric. In *Proceedings of the 3rd annual ACM conference on Research and development in information retrieval*. Citeseer, 256–267.
- [13] Luca Busin and Stefano Mizzaro. 2013. Axiometrics: An axiomatic approach to information retrieval effectiveness metrics. In *Proceedings of the 2013 Conference on the Theory of Information Retrieval*. 22–29.
- [14] Ben Carterette. 2011. System effectiveness, user models, and user utility: a conceptual framework for investigation. In *Proceedings of the 34th international ACM SIGIR conference on Research and development in information retrieval*. 903–912.
- [15] Cyril Cleverdon. 1967. The Cranfield tests on index language devices. In *Aslib proceedings*, Vol. 19. MCB UP Ltd, 173–194.
- [16] Stéphane Clinchant and Eric Gaussier. 2011. Is document frequency important for PRF?. In *Conference on the theory of information retrieval*. Springer, 89–100.
- [17] Stéphane Clinchant and Eric Gaussier. 2013. A theoretical analysis of pseudo-relevance feedback models. In *Proceedings of the 2013 Conference on the Theory of Information Retrieval*. 6–13.
- [18] Byron E Dom. 2012. An information-theoretic external cluster-validity measure. *arXiv preprint arXiv:1301.0565* (2012).
- [19] Hui Fang. 2007. *An axiomatic approach to information retrieval*. Technical Report.
- [20] Hui Fang, Tao Tao, and ChengXiang Zhai. 2004. A formal study of information retrieval heuristics. In *Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval*. 49–56.
- [21] Hui Fang, Tao Tao, and ChengXiang Zhai. 2011. Diagnostic evaluation of information retrieval models. *ACM Transactions on Information Systems (TOIS)* 29, 2 (2011), 1–42.
- [22] Hui Fang and ChengXiang Zhai. 2005. An exploration of axiomatic approaches to information retrieval. In *Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval*. 480–487.
- [23] Hui Fang and ChengXiang Zhai. 2006. Semantic term matching in axiomatic approaches to information retrieval. In *Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval*. 115–122.
- [24] Marco Ferrante, Nicola Ferro, and Norbert Fuhr. 2021. Towards Meaningful Statements in IR Evaluation: Mapping Evaluation Measures to Interval Scales. *arXiv preprint arXiv:2101.02668* (2021).
- [25] Marco Ferrante, Nicola Ferro, and Maria Maistro. 2015. Towards a formal framework for utility-oriented measurements of retrieval effectiveness. In *Proceedings of the 2015 International Conference on the Theory of Information Retrieval*. 21–30.
- [26] Marco Ferrante, Nicola Ferro, and Silvia Pontarollo. 2018. A general theory of IR evaluation measures. *IEEE Transactions on Knowledge and Data Engineering* 31, 3 (2018), 409–422.
- [27] Norbert Fuhr. 2021. Proof by experimentation? towards better IR research. In *ACM SIGIR Forum*, Vol. 54. ACM New York, NY, USA, 1–4.
- [28] George Grätzer. 2002. *General lattice theory*. Springer Science & Business Media.
- [29] Matthias Hagen, Michael Völske, Steve Göring, and Benno Stein. 2016. Axiomatic result re-ranking. In *Proceedings of the 25th ACM International on Conference on Information and Knowledge Management*. 721–730.
- [30] Theodorus Wilhelmus Charles Huibers. 1996. *An axiomatic theory for information retrieval*. Ph. D. Dissertation.
- [31] Maryam Karimzadehgan and ChengXiang Zhai. 2012. Axiomatic analysis of translation language model for information retrieval. In *European Conference on Information Retrieval*. Springer, 268–280.
- [32] Jaana Kekäläinen and Kallervo Järvelin. 2002. Using graded relevance assessments in IR evaluation. *Journal of the American Society for Information Science and Technology* 53, 13 (2002), 1120–1129.
- [33] David Krantz, Duncan Luce, Patrick Suppes, and Amos Tversky. 1971. Foundations of measurement, Vol. I: Additive and polynomial representations. (1971).
- [34] David H Krantz. 1989. Foundations of Measurement. Vol. II. Geometrical, Threshold and Probabilistic Representations. (1989).
- [35] Eddy Maddalena and Stefano Mizzaro. 2014. Axiometrics: Axioms of Information Retrieval Effectiveness Metrics.. In *EVIA@ NTCIR*.
- [36] Marina Meilä. 2007. Comparing clusterings—an information based distance. *Journal of multivariate analysis* 98, 5 (2007), 873–895.
- [37] Alistair Moffat. 2013. Seven numeric properties of effectiveness metrics. In *Asia Information Retrieval Symposium*. Springer, 1–12.
- [38] Alistair Moffat, Paul Thomas, and Falk Scholer. 2013. Users versus models: What observation tells us about effectiveness metrics. In *Proceedings of the 22nd ACM international conference on Information & Knowledge Management*. 659–668.
- [39] Alistair Moffat and Justin Zobel. 2008. Rank-biased precision for measurement of retrieval effectiveness. *ACM Transactions on Information Systems (TOIS)* 27, 1 (2008), 1–27.
- [40] Ali Montazerlghaem, Hamed Zamani, and Azadeh Shakery. 2016. Axiomatic analysis for improving the log-logistic feedback model. In *Proceedings of the 39th International ACM SIGIR conference on Research and Development in Information Retrieval*. 765–768.
- [41] Razieh Rahimi, Ali Montazerlghaem, and Azadeh Shakery. 2020. An axiomatic approach to corpus-based cross-language information retrieval. *Information Retrieval Journal* 23, 3 (2020), 191–215.
- [42] Fred S Roberts. 1985. Measurement theory. *Encyclopedia of Mathematics and its applications* 7 (1985).
- [43] Andrew Rosenberg and Julia Hirschberg. 2007. V-measure: A conditional entropy-based external cluster evaluation measure. In *Proceedings of the 2007 joint conference on empirical methods in natural language processing and computational natural language learning (EMNLP-CoNLL)*. 410–420.
- [44] Corby Rosset, Bhaskar Mitra, Chenyan Xiong, Nick Craswell, Xia Song, and Saurabh Tiwary. 2019. An axiomatic approach to regularizing neural ranking models. In *Proceedings of the 42nd international ACM SIGIR conference on research and development in information retrieval*. 981–984.
- [45] Gian-Carlo Rota. 1971. On the combinatorics of the Euler characteristic. In *Studies in Pure Mathematics (Presented to Richard Rado)*. Academic Press London, 221–233.
- [46] Tetsuya Sakai. 2021. On Fuhr’s guideline for IR evaluation. In *ACM SIGIR Forum*, Vol. 54. ACM New York, NY, USA, 1–8.
- [47] Tetsuya Sakai and Noriko Kando. 2008. On information retrieval metrics designed for evaluation with incomplete relevance assessments. *Information Retrieval* 11, 5 (2008), 447–470.
- [48] Fabrizio Sebastiani. 2015. An axiomatically derived measure for the evaluation of classification algorithms. In *Proceedings of the 2015 international conference on the theory of information retrieval*. 11–20.
- [49] Fabrizio Sebastiani. 2020. Evaluation measures for quantification: An axiomatic approach. *Information Retrieval Journal* 23, 3 (2020), 255–288.
- [50] Marina Sokolova. 2006. Assessing invariance properties of evaluation measures. In *Proceedings of the Workshop on Testing of Deployable Learning and Decision Systems, the 19th Neural Information Processing Systems Conference (NIPS 2006)*.
- [51] John A Swets. 1963. Information retrieval systems. *Science* 141, 3577 (1963), 245–250.
- [52] Cornelis Joost Van Rijsbergen. 1974. Foundation of evaluation. *Journal of documentation* 30, 4 (1974), 365–373.
- [53] Michael Völske, Alexander Bondarenko, Maik Fröbe, Benno Stein, Jaspreet Singh, Matthias Hagen, and Avishek Anand. 2021. Towards Axiomatic Explanations for Neural Ranking Models. In *Proceedings of the 2021 ACM SIGIR International Conference on Theory of Information Retrieval*. 13–22.
- [54] YY Yao. 1995. Measuring retrieval effectiveness based on user preference of documents. *Journal of the American Society for Information science* 46, 2 (1995), 133–145.
- [55] ChengXiang Zhai and Hui Fang. 2013. Axiomatic analysis and optimization of information retrieval models. In *Proceedings of the 2013 Conference on the Theory of Information Retrieval*. 3–3.

A FORMAL PROOFS

PROOF. [Proposition 1]:

It will be seen that $\{\max\{\hat{r}_1, \hat{s}_1\}, \dots, \max\{\hat{r}_N, \hat{s}_N\}\}$ is the lower upper bound of \hat{r} and \hat{s} .

First, it is an upper bound of \hat{r} and \hat{s} since $\hat{r}_i \leq \max\{\hat{r}_i, \hat{s}_i\}$, $\hat{s}_i \leq \max\{\hat{r}_i, \hat{s}_i\}$, $\forall i \in \{1, \dots, N\}$, and these conditions fulfil the definition of the **replacement** [set-based] partial ordering.

Now, it will be seen that it is the lower upper bound. Let $\hat{t} = \{\hat{t}_1, \dots, \hat{t}_N\} \in R(N)$, such that $\hat{r} \leq \hat{t}$ and $\hat{s} \leq \hat{t}$. Then, $|\{i : \hat{r}_i \geq a_j\}| \leq |\{i : \hat{t}_i \geq a_j\}|$, $\forall j \in \{0, \dots, c\}$ and $|\{i : \hat{s}_i \geq a_j\}| \leq |\{i : \hat{t}_i \geq a_j\}|$, $\forall j \in \{0, \dots, c\}$. Therefore, $|\{i : \max\{\hat{r}_i, \hat{s}_i\} \geq a_j\}| \leq |\{i : \hat{t}_i \geq a_j\}|$, $\forall j \in \{0, \dots, c\}$.

The other equality can be seen by duality. \square

PROOF. [Proposition 2]:

The **replacement** [set-based] partial order is grounded on the *replacement* operation. It will be seen that one system run, $\hat{s} \in R(N)$, covers another, $\hat{r} \in R(N)$, if they differ only in the *replacement* of a document with the relevance degree immediately consecutive in *REL*. Formally:

$$\hat{r} < \hat{s} \iff \exists k : \hat{r}_k = a_i, \hat{s}_k = a_{i+1} \wedge \hat{r}_j = \hat{s}_j, \forall j \neq k$$

This is true, since if $\exists k : \hat{r}_k = a_i, \hat{s}_k = a_{i+1} \wedge \hat{r}_j = \hat{s}_j, \forall j \neq k$, then \hat{s} has more *cumulative mass of relevance* than \hat{r} for every relevance degree, which implies that $\hat{r} \leq \hat{s}$. In addition, there cannot be a different judged run between \hat{r} and \hat{s} since they differ in only a document that has been classified with the consecutive relevance degree.

The aim is to apply Theorem 1 to conclude that $(R(N), \wedge, \vee, \leq)$ is a distributive lattice. Thus, it is necessary to see that $R(N)$ does not contain a sublattice equal to N_5 or M_3 .

Suppose that $R(N)$ contains a sublattice N_5 as depicted in Figure 1, where $a = \hat{u}$, $b = \hat{r}^1$, $c = \hat{r}^2$, $d = \hat{s}$ and $e = \hat{v}$. It can be postulated that $\hat{u} = \{a_i, a_j, a_k\}$, where $a_i, a_j, a_k \in REL$. Perhaps, \hat{u} may contain more relevance degrees; however, they do not participate in the demonstration. Thus, this hypothesis can be assumed to hold.

According to the covering relation, \hat{r}^1 and \hat{s} have to increase one relevance degree of \hat{u} , being consecutive in REL . Notice that they cannot increase the same relevance degree, since they are placed in separate paths, in the Hasse diagram. Thus, it can be assumed that $\hat{r}^1 = \{a_{i+1}, a_j, a_k\}$ and $\hat{s} = \{a_i, a_{j+1}, a_k\}$.

Then, \hat{r}^2 has to increase one of the relevance degrees of \hat{r}^1 . There are three possibilities, which will be analysed in the following:

- (1) Suppose that a_{i+1} has been increased, i.e., $\hat{r}^2 = \{a_{i+2}, a_j, a_k\}$. Then, the configuration of \hat{v} would contain a_{i+2} or a_{i+3} , depending on if there is an increase in that relevance degree. However, both cases are not possible when the covering relation is applied to \hat{s} since $\hat{s} = \{a_i, a_{j+1}, a_k\}$, and a_i can increase only one unity in REL .
- (2) Suppose that a_k has been increased, i.e., $\hat{r}^2 = \{a_{i+1}, a_j, a_{k+1}\}$. Then, this configuration is not compatible with \hat{s} since it can increase only a_i or a_k but not both.
- (3) Suppose that a_j has been increased, i.e., $\hat{r}^2 = \{a_{i+1}, a_{j+1}, a_k\}$. Then, \hat{s} has to increase a_i and any increase in \hat{r}^2 will not be compatible with \hat{s} .

As every case is not possible, $R(N)$ does not contain any N_5 .

Now, suppose that $R(N)$ contains a sublattice M_3 as depicted in Figure 1, where $a = \hat{u}$, $b = \hat{r}$, $c = \hat{s}$, $d = \hat{t}$ and $e = \hat{v}$. It can be postulated that $\hat{u} = \{a_i, a_j, a_k\}$, where $a_i, a_j, a_k \in REL$. Perhaps, \hat{u} may contain more relevance degrees; however, they do not take part in the demonstration. Thus, this hypothesis can be assumed to hold.

According to the covering relation, \hat{r} , \hat{s} and \hat{t} have to increase one relevance degree of \hat{u} , being consecutive in REL . Note that they cannot increase the same relevance degree since they are placed in separate paths in the Hasse diagram. Thus, it can be assumed that $\hat{r} = \{a_{i+1}, a_j, a_k\}$, $\hat{s} = \{a_i, a_{j+1}, a_k\}$ and $\hat{t} = \{a_i, a_j, a_{k+1}\}$.

Then, \hat{v} will be obtained by increasing one relevance degree of \hat{r} , \hat{s} and \hat{t} . For example, if \hat{r} increases a_j , then $\hat{v} = \{a_{i+1}, a_{j+1}, a_k\}$. This configuration is compatible with \hat{s} , when a_i is increased. However, \hat{v} is not compatible with \hat{t} since it can only increase a_i or a_j but not both. Similar reasoning shows that the rest of the cases are not possible.

Thus, $R(N)$ cannot contain M_3 as a sublattice. Then, according to Theorem 1, $R(N)$ is a distributive lattice. \square

PROOF. [Proposition 3]:

In the demonstration of Proposition 2, it is shown that the covering relation of the **replacement** [set-based] ordering is as follows: given a pair of system runs, $\hat{r}, \hat{s} \in R(N)$ with $\hat{r} \neq \hat{s}$, then $\hat{r} \prec \hat{s} \Leftrightarrow \exists k : \hat{r}_k = a_i, \hat{s}_k = a_{i+1}$ and $\hat{r}_j = \hat{s}_j, \forall j \neq k$.

The join-irreducible elements are those which only have one descendant in the Hasse diagram. It will be seen that these elements are those which only have two different relevance degrees in the configuration, and one of them is a_0 .

Let \hat{r} be a judged run with at least two documents classified with relevance degrees different from a_0 . Thus, $\hat{r} = \{a_i, a_j\} \cup \hat{r}'$, where $i, j \neq 0$ and \hat{r}' could be any subset of judged documents, even the empty set. Then, according to the covering relation, \hat{r} has at least two descendants in the Hasse diagram. They are: $\{a_{i-1}, a_j\} \cup \hat{r}'$ and $\{a_i, a_{j-1}\} \cup \hat{r}'$, thus \hat{r} cannot be a join-irreducible element.

Finally, the judged runs listed in the statement of this Proposition are the only ones which have one or two relevance degrees, where one of them is

a_0 . It will be seen that these elements only have one descendant in the Hasse diagram. Let $\hat{r} = \{a_i, \dots, a_i, a_0, \dots, a_0\}$ be one of these elements, then the only descendant in the Hasse diagram is $\{a_i, \dots, a_i, a_{i-1}, a_0, \dots, a_0\}$ since it only can have one relevance degree different and consecutive in REL . \square

PROOF. [Proposition 4]:

Given two system runs $\hat{r}, \hat{s} \in R(N)$, it will be seen that $gP(\hat{r} \vee \hat{s}) + gP(\hat{r} \wedge \hat{s}) = gP(\hat{r}) + gP(\hat{s})$. Taking into account the definition of the meet and the join in Proposition 1,

$$\begin{aligned} gP(\hat{r} \vee \hat{s}) + gP(\hat{r} \wedge \hat{s}) &= \frac{1}{N} \cdot \sum_{i=1}^N \frac{\max\{g(\hat{r}_i), g(\hat{s}_i)\}}{g(a_c)} + \frac{1}{N} \cdot \sum_{i=1}^N \frac{\min\{g(\hat{r}_i), g(\hat{s}_i)\}}{g(a_c)} = \\ &= \frac{1}{N} \cdot \sum_{i=1}^N \frac{\max\{g(\hat{r}_i), g(\hat{s}_i)\} + \min\{g(\hat{r}_i), g(\hat{s}_i)\}}{g(a_c)} = \frac{1}{N} \cdot \sum_{i=1}^N \frac{g(\hat{r}_i) + g(\hat{s}_i)}{g(a_c)} = \\ &= \frac{1}{N} \cdot \sum_{i=1}^N \frac{g(\hat{r}_i)}{g(a_c)} + \frac{1}{N} \cdot \sum_{i=1}^N \frac{g(\hat{s}_i)}{g(a_c)} = gP(\hat{r}) + gP(\hat{s}) \end{aligned}$$

The same demonstration is valid for the generalized recall, where N should be changed by RB . \square

PROOF. [Proposition 5]: The demonstration is quite similar to the proof of Proposition 1. It will be seen that $(\max\{\hat{r}_1, \hat{s}_1\}, \dots, \max\{\hat{r}_N, \hat{s}_N\})$ is the lower upper bound of \hat{r} and \hat{s} .

First, it is an upper bound of \hat{r} and \hat{s} since $\hat{r}[i] \leq \max\{\hat{r}[i], \hat{s}[i]\}$, $\hat{s}[i] \leq \max\{\hat{r}[i], \hat{s}[i]\}$, $\forall i \in \{1, \dots, N\}$, and these conditions fulfil the definition of the **replacement** [rank-based] partial ordering.

Now, it will be seen that it is the lower upper bound. Let $\hat{t} = (\hat{t}[1], \dots, \hat{t}[N]) \in R(N)$, such that $\hat{r} \leq \hat{t}$ and $\hat{s} \leq \hat{t}$. Then, $|\{i : \hat{r}[i] \geq a_j\}| \leq |\{i : \hat{t}[i] \geq a_j\}|, \forall j \in \{0, \dots, c\}$ and $|\{i : \hat{s}[i] \geq a_j\}| \leq |\{i : \hat{t}[i] \geq a_j\}|, \forall j \in \{0, \dots, c\}$. Therefore $|\{i : \max\{\hat{r}[i], \hat{s}[i]\} \geq a_j\}| \leq |\{i : \hat{t}[i] \geq a_j\}|, \forall j \in \{0, \dots, c\}$.

The other equality can be seen by duality. \square

PROOF. [Proposition 7]: The demonstration is quite similar to the proof of Proposition 3.

In the demonstration of the Proposition 6, it is shown that the covering relation of the **replacement** [rank-based] ordering is as follows: given a pair of system runs, $\hat{r}, \hat{s} \in R(N)$ with $\hat{r} \neq \hat{s}$, then $\hat{r} \prec \hat{s} \Leftrightarrow \exists k : \hat{r}[k] = a_i, \hat{s}[k] = a_{i+1}$ and $\hat{r}[j] = \hat{s}[j], \forall j \neq k$.

The join-irreducible elements are those which only have one descendant in the Hasse diagram. It will be seen that these elements are those which only have one component different from a_0 .

Let \hat{r} be a judged run with at least two components with relevance degrees different from a_0 . Thus, $\hat{r} = (\dots, a_i, \dots, a_j, \dots)$, where $i, j \neq 0$. Then, according to the covering relation, \hat{r} has at least two descendants in the Hasse diagram. They are: $(\dots, a_{i-1}, \dots, a_j, \dots)$ and $(\dots, a_i, \dots, a_{j-1}, \dots)$, thus \hat{r} cannot be a join-irreducible element.

Finally, the judged runs listed in the statement of this Proposition are the ones which only have one component different from a_0 . It will be seen that these elements only have one descendant in the Hasse diagram. Let $\hat{r} = (a_0, \dots, a_i, \dots, a_0)$ be one of these elements, then the only descendant in the Hasse diagram is $(a_0, \dots, a_{i-1}, \dots, a_0)$ since it only can have one relevance degree different and consecutive in REL . \square

The proofs of Propositions 6 and 8 are quite similar to the proofs of Propositions 2 and 4, respectively, which have been omitted for space limitations.