

An information measure for comparing top k lists

James H. Collier[†], and Arun S. Konagurthu^{†*}

[†] Clayton School of Information Technology, Faculty of Information Technology, Monash University, VIC 3800 Australia

* Corresponding author. (E-mail: arun.konagurthu@monash.edu).

Abstract—Comparing the top k elements between two or more ranked results is a common task in many contexts and settings. A few measures have been proposed to compare top k lists with attractive mathematical properties, but they face a number of pitfalls and shortcomings in practice. This work introduces a new measure to compare any two top k lists based on measuring the information these lists convey. Our method investigates the compressibility of the lists, and the length of the message to encode losslessly the lists gives a natural and robust measure of their variability. This information-theoretic measure objectively reconciles all the main considerations that arise when measuring (dis-)similarity between lists: the extent of their non-overlapping elements; the amount of disarray among overlapping elements; the measurement of displacement of actual ranks (positions) of their overlapping elements. We demonstrate that our measure is intuitively simple and superior to other commonly used measures. To the best of our knowledge, this is the first attempt to address the problem using information compression as its basis.

I. INTRODUCTION

Ranked results are handled in diverse settings, from the web page results of search engines to genes in differential gene co-expression experiments. A routine task that follows is the assessment of variability between top k ranked elements between two or more related lists.

Comparing top k lists has received much attention over the past decade. Among the most cited work on this topic is that of Fagin and colleagues [1]. They proposed an easy-to-compute metric built on Spearman's foot rule [2]. Formally, if π_1 and π_2 define two permutations from the symmetric group S_n of all permutations of n elements, Spearman's foot rule gives the L_1 distance between the ranks of corresponding elements in the two permutations as: $L_1(\pi_1, \pi_2) = \sum_{i=1}^n |\pi_1(i) - \pi_2(i)|$, where any $\pi_1(i)$ or $\pi_2(i)$ is the position (rank) of the i th element in the permutation, given some total ordering of n elements. Fagin *et al.* extend this metric to comparing two top k lists in presence of non-overlapping elements (*i.e.*, elements that are in one list but not in the other). This is achieved by fixing the contribution, to the distance, of the non-overlapping elements to a value greater than k , typically $(k+1)$. Formally, the extended metric for two top k lists τ_1 and τ_2 is defined as $L_1(\tau_1, \tau_2) = 2(k - |\tau_1 \cap \tau_2|)(k+1) + \sum_{i \in \tau_1 \cap \tau_2} |\tau_1(i) - \tau_2(i)| - \sum_{i \in \tau_1 - \tau_2} \tau_1(i) - \sum_{i \in \tau_2 - \tau_1} \tau_2(i)$ where $\tau_1 \cap \tau_2$ is the set of elements that overlap between the two lists, $|\tau_1 \cap \tau_2|$ denotes the number of overlapping

elements, $\tau_1 - \tau_2$ gives the non-overlapping elements in τ_1 , and $\tau_2 - \tau_1$ gives those in τ_2 .

Although this measure can be shown to have good mathematical properties, in practice, it has crucial limitations. Mainly, it can be seen from the formulation that the term $2(k - |\tau_1 \cap \tau_2|)(k+1)$ grows quadratically for increasing values of k and decreasing proportion of overlapping elements. In fact, in many applications requiring comparison of top k lists (*e.g.*, web search results), non-overlapping elements form a significant proportion of the lists. Furthermore, this metric is insensitive to the absolute ranks of the overlapping elements in the respective lists; when computing the L_1 distance, the overlapping elements are re-ranked, and hence ignore the displacement of these elements when comparing two lists.

Another commonly used metric is based on Kendall tau distance [3], or, colloquially, the *bubble-sort distance*, since it measures the number of adjacent transpositions required to convert (*i.e.*, sort) one permutation to another. Formally, for any two permutations π_1 and π_2 , Kendall tau distance is defined (using the same notations as above) as $K(\pi_1, \pi_2) = \sum_{\forall 1 \leq i < j \leq n} \kappa_{i,j}(\pi_1, \pi_2)$, where $\kappa_{i,j}(\pi_1, \pi_2) = 0$ if $\pi_1(i) < \pi_1(j)$ and $\pi_2(i) < \pi_2(j)$, or $\kappa_{i,j}(\pi_1, \pi_2) = 1$ otherwise. Extending this idea, the following cost function was proposed to compare two top k lists [4]: $K(\tau_1, \tau_2) = (k - |\tau_1 \cap \tau_2|)((2+p)k - p|\tau_1 \cap \tau_2| + 1 - p) + \sum_{i \in \tau_1 \cap \tau_2} \kappa_{i,j}(\tau_1, \tau_2) - \sum_{i \in \tau_1 - \tau_2} \tau_1(i) - \sum_{i \in \tau_2 - \tau_1} \tau_2(i)$ where, p is a tunable penalty parameter to account for the transposition distance between non-overlapping elements in τ_1 and τ_2 . However, it is easy to see that this metric is also sensitive to the size of non-overlapping elements in the two lists, in addition to the choice of penalty parameter p .

Other measures have been proposed on specialized applications [5], [6], [7], [8], [9], [10]. Noteworthy among these is the Canberra distance [11] between top k lists [10]. This distance is a weighted variant of Spearman's L_1 distance, and ensures that the displacement of elements with higher ranks is penalized more than those with lower ranks.

Our results: Here, we introduce a new information measure to compare any two top k lists. We build our method on the statistical framework of minimum length encoding introduced by Chris Wallace [12], [13]. Our method investigates the compressibility of top k lists. It is intuitive to see that closely related lists have more information in common (and hence more compressible) than the lists

that are poorly related. Thus, the *length* of the losslessly compressed message gives a natural and rigorous measure to estimate the variability between two lists. Unlike previous work, this measure achieves an *objective* trade-off between conflicting criteria, measuring the variability between lists. Mainly, these include: (1) the measurement of dissimilarity, (2) the measurement of disarray of its overlapping elements, and (3) the displacement of the positions (ranks) of these elements.

We note that measuring the *true* information content of any data is incomputable. This follows from the fact that Solomonoff-Kolmogorov-Chaitin Complexity [14], [15], [16] is undecidable. However, effective and efficient statistical models for data compression provide reasonable upper bounds (*i.e.*, estimates) of true information content. Further, our paper provides an approach to estimating the information content in any given pair of top k lists. To keep this approach general, our models of compression use simple priors. However it is important to note that this information theoretic framework can be adapted to individual contexts by accommodating prior knowledge about rankings in those settings.

Organization of the paper: This paper is organized as follows. Section II introduces our information measure formally and describes some interesting mathematical properties. Section III explains the practical details involved in estimating the information content of two lists. Section IV presents the experimental results of comparing our information measure with other popular distance metrics on ranked lists.

II. INFORMATION MEASURE FOR COMPARING RANKED LISTS

The mathematical underpinning of our information measure to compare any two top k lists is established here. For details of how this information measure defined in this section is realized in practice, see the following Section III.

Definition 1. (*Shannon's information content of an outcome*)

Information content of an outcome E whose probability is $P(E)$ is given by $I(E) = -\log(P(E))$.¹

We note that $I(E)$ corresponds to the (theoretical) lower bound on the length of the optimal code required to *losslessly* encode the outcome E . This result comes from the Shannon's seminal work on the mathematical theory of communication [17].

Lemma 1. (*Measure of Information between two top k lists*)
For two top k lists, τ_1 and τ_2 , the total amount of information contained in them is $I(\tau_1, \tau_2) = I(\tau_1) + I(\tau_2|\tau_1)$

¹Base of the logarithm gives the information measure its units. \log_2 yields information measured in bits. \ln gives nits or nats, and \log_{10} , dits or hartleys.

Proof: Using the product rule of probability, the joint probability of the two top k lists, $\Pr(\tau_1, \tau_2)$ is the product of the probability of the first list, $\Pr(\tau_1)$, and the probability of the second list conditioned on the first, $\Pr(\tau_2|\tau_1)$:

$$\Pr(\tau_1, \tau_2) = \Pr(\tau_1) \times \Pr(\tau_2|\tau_1)$$

Taking negative logarithm on both sides and applying Shannon's insight in Definition 1 to these probabilities, gives the identity: $I(\tau_1, \tau_2) = I(\tau_1) + I(\tau_2|\tau_1)$ ■

Lemma 2. $I(\tau_1, \tau_2) \leq I(\tau_1) + I(\tau_2)$

Proof: When the two top k lists are independent of each other

$$\Pr(\tau_1, \tau_2) = \Pr(\tau_1) \times \Pr(\tau_2|\tau_1) = \Pr(\tau_1) \times \Pr(\tau_2)$$

This implies that $\Pr(\tau_1, \tau_2) \geq \Pr(\tau_1) \times \Pr(\tau_2)$. Translating this into information terms by taking negative logarithm on both sides, we get $I(\tau_1, \tau_2) \leq I(\tau_1) + I(\tau_2)$. ■

More plainly, if τ_1 and τ_2 are independent of each other, that is the knowledge of one list does not inform the contents of the other list, the joint information content in these lists is the sum of the information content in each of the lists taken separately, *i.e.*, $I(\tau_1) + I(\tau_2)$. We use the term $NULL(\tau_1, \tau_2)$ in this work to define this upper bound on the joint information in the two top k lists.

Lemma 3. *Given three top k lists, τ_1 , τ_2 and τ_3 ,*

$$I(\tau_1, \tau_2) - I(\tau_1, \tau_3) = \log \left(\frac{\Pr(\tau_3|\tau_1)}{\Pr(\tau_2|\tau_1)} \right)$$

Proof: Using Lemma 1

$$\begin{aligned} I(\tau_1, \tau_2) &= I(\tau_1) + I(\tau_2|\tau_1) \quad \text{and} \\ I(\tau_1, \tau_3) &= I(\tau_1) + I(\tau_3|\tau_1). \end{aligned}$$

Subtracting the two terms,

$$\begin{aligned} I(\tau_1, \tau_2) - I(\tau_1, \tau_3) &= I(\tau_1) + I(\tau_2|\tau_1) - (I(\tau_1) + I(\tau_3|\tau_1)) \\ &= I(\tau_2|\tau_1) - I(\tau_3|\tau_1) \\ &= -\log(\Pr(\tau_2|\tau_1)) + \log(\Pr(\tau_3|\tau_1)) \\ &= \log \left(\frac{\Pr(\tau_3|\tau_1)}{\Pr(\tau_2|\tau_1)} \right). \end{aligned}$$

In other words, the difference above gives the log-odds conditional probability (or posterior) ratio between the lists being compared. This establishes a rigorous statistical framework to compare ranked lists.

A corollary of this property is that we can define the information divergence between any two given top k lists by treating τ_3 as a separate list which happens to be an identical copy of τ_1 as defined below.

Definition 2. (*Information divergence or cost*)

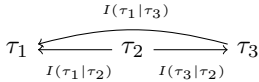
In this work, we measure the information distance between

two top k lists as $I(\tau_1, \tau_2) - I(\tau_1, (\tau_3 \equiv \tau_1)) = I(\tau_2|\tau_1) - I((\tau_3 \equiv \tau_1)|\tau_1)$.²

Since the information divergence defined above is a function of conditional information terms on the right hand side, we analyse below the metrical properties of conditional information between top k lists.

Property 1. (*Directed acyclic triangular inequality of conditional information*)

For three top k lists, τ_1 , τ_2 , and τ_3 , we have:

$$I(\tau_1|\tau_2) \leq I(\tau_1|\tau_3) + I(\tau_3|\tau_2)$$


Proof: This follows by expanding the joint information in the three lists as follow:

$$\begin{aligned} I(\tau_1, \tau_2, \tau_3) &= I(\tau_3) + I(\tau_1, \tau_2|\tau_3) = I(\tau_2) + I(\tau_1, \tau_3|\tau_2) \\ &= I(\tau_3) + I(\tau_1|\tau_3) + I(\tau_2|\tau_1, \tau_3) \\ &= I(\tau_2) + I(\tau_1, \tau_3|\tau_2) \\ &= I(\tau_3) + I(\tau_1|\tau_3) + I(\tau_2|\tau_3) \\ &\geq I(\tau_2) + I(\tau_1, \tau_3|\tau_2) \end{aligned}$$

Rearranging terms, we get:

$$\begin{aligned} &(I(\tau_3) + I(\tau_2|\tau_3)) + I(\tau_1|\tau_3) \geq I(\tau_2) + I(\tau_1, \tau_3|\tau_2) \\ \implies &I(\tau_2, \tau_3) + I(\tau_1|\tau_3) \geq I(\tau_2) + I(\tau_1, \tau_3|\tau_2) \\ \implies &(I(\tau_2, \tau_3) - I(\tau_2)) + I(\tau_1|\tau_3) \geq I(\tau_1, \tau_3|\tau_2) \\ \implies &I(\tau_3|\tau_2) + I(\tau_1|\tau_3) \geq I(\tau_1, \tau_3|\tau_2) \\ \implies &I(\tau_3|\tau_2) + I(\tau_1|\tau_3) \geq I(\tau_1|\tau_2) + I(\tau_3|\tau_1, \tau_2) \\ \implies &I(\tau_3|\tau_2) + I(\tau_1|\tau_3) \geq I(\tau_1|\tau_2) \end{aligned}$$

■

Property 2. (*Conditional Information is not symmetric*)

$$I(\tau_1|\tau_2) \neq I(\tau_2|\tau_1)$$

Proof: From the product rule of probability [18], we have

$$\Pr(\tau_1, \tau_2) = \Pr(\tau_1) \Pr(\tau_2|\tau_1) = \Pr(\tau_2) \Pr(\tau_1|\tau_2).$$

Applying Definition 1 we get

$$I(\tau_1, \tau_2) = I(\tau_1) + I(\tau_2|\tau_1) = I(\tau_2) + I(\tau_1|\tau_2)$$

Rearranging terms above

$$I(\tau_1|\tau_2) = I(\tau_2|\tau_1) + I(\tau_2) - I(\tau_1),$$

² $I((\tau_3 \equiv \tau_1), \tau_1)$ should not be confused with $I(\tau_1)$ as the former measures the joint information in two *separate* lists that happen to be identical. For $I(\tau_3 \equiv \tau_1, \tau_1) = I(\tau_1) + I((\tau_3 \equiv \tau_1)|\tau_1) = I(\tau_1)$ implies that the conditional probability $I((\tau_3 \equiv \tau_1)|\tau_1) = 0$, which is same as saying $\Pr((\tau_3 \equiv \tau_1)|\tau_1) = 1$, suggesting absolute certainty that τ_3 is identical to τ_1 ; this will be incorrect.

proving the asymmetry between $I(\tau_1|\tau_2)$ and $I(\tau_2|\tau_1)$. ■

Property 3. (*Near-coincidence of conditional information*)
For any two separate yet identical top k lists τ_1 and $\tau_2 \equiv \tau_1$ $I((\tau_2 \equiv \tau_1)|\tau_1) = \epsilon$, where ϵ is some small constant.

Proof: This follows because $I(\tau_1, \tau_2) = I(\tau_1) + I(\tau_1|\tau_2)$. Since τ_2 is same as τ_1 , the additional conditional information required to state a list which is an identical copy of another list is a small constant (in number of bits/nits required to transmit this identity). ■

These properties suggest that the information measure defined here possesses near-metrical properties, which are useful to compare top k lists.

III. REALIZING THE INFORMATION MEASURE IN PRACTICE

This section will describe an approach to realize an information theoretic measure to quantify the variability of two top k lists. To make this measure intuitively understood, this information measure can be rationalized as a communication process between an imaginary pair of transmitter (Alice) and receiver (Bob).

Imagine Alice has access to two top k lists τ_1 and τ_2 . Her goal is to communicate the information in both these lists to Bob *losslessly* – exactly as she sees it. To achieve this, Alice constructs a two-part message. In the first, she will transmit τ_1 taking $I(\tau_1)$ bits. In the second, she uses the commonality (if any) between the two lists so that τ_2 can be transmitted more concisely; this takes $I(\tau_2|\tau_1)$ bits.

In this information theoretic framework the measure of (dis-)similarity between two top k lists is the *total length of this two-part message*: $I(\tau_1) + I(\tau_2|\tau_1)$. It is easy to see that if $\tau_2 \equiv \tau_1$, the second part is extremely concise. On the other hand, if τ_2 is completely unrelated to τ_1 , then the amount of information to transmit the second list given the first, $I(\tau_2|\tau_1)$, cannot be shorter than $I(\tau_2)$.

To achieve lossless transmission of the two lists between Alice and Bob, the following pieces of information are necessary:

- 1) The size $k = |\tau_1| = |\tau_2|$ of these top k lists.
- 2) The elements in τ_1 , in the order they appear.
- 3) The overlapping elements between τ_1 and τ_2 .
- 4) The ranks of the overlapping elements in τ_2 .
- 5) The permutation of overlapping elements in τ_2 with respect to the ordering defined by τ_1 .
- 6) The non-overlapping elements in τ_2 in the order they appear in that list.

In transmitting the above information, two distinct cases have to be considered:

Case 1: When the domain of elements that are being ranked is *known*, of which τ_1 and τ_2 are (partial) top k lists. For instance, it is common in differential gene expression studies that total domain of

genes and their labels (identifiers) is known, and a set of top 50 differentially expressed genes between two experiments are considered.

Case 2: Conversely, when the domain of ranked elements remains *unknown*. For instance, in search results from popular web search engines, while we see the top search results, the number of pages each search engine indexes is variable (and unknown) and is often smaller than the universe of pages available on the internet.

The remaining part of this section, we handle these two cases separately and describe the encoding schemes to transmit each of the enumerated pieces of information.

Case 1: When the domain of elements is known

Here we assume that the size (N) of the domain is known along with the labels (or identifiers) of elements in it.

Step 1: Transmitting the size of the top k lists: The size of $k \leq N$ is transmitted as an integer code. Since both Alice and Bob know that the top k lists come from a domain of N elements, a simple encoding of k takes $\log(N)$ bits, assuming an uniform distribution over the choices of k in the range $1 \leq k \leq N$. We note that more sophisticated encodings can be conceived if their is a prior belief that the distribution of k is non-uniform.

Step 2: Transmitting τ_1 : Transmitting the information in τ_1 can be achieved by communicating, over an integer code, the lexicographic index associated with τ_1 in some (mutually agreed) lexicographic ordering of the k -permutations of N elements. Since both Alice and Bob know the domain from which the ranking was generated, the lexicographic ordering of k -permutations can be treated as a part of the code book of communication.

Step 3: Transmitting overlapping elements between τ_1 and τ_2 : At this stage Bob already knows τ_1 . To nominate the overlapping elements, that is, the intersection between the two top k lists, a bit mask b_1 is defined where the set bits indicate the positions in τ_1 where the overlapping elements reside. Transmission complexity of stating the intersection between τ_1 and τ_2 is same as the complexity of this bit mask. An efficient encoding scheme to transmit this bit mask, assuming no prior knowledge about the distribution of the set bits, would be using an adaptive code over a binomial distribution.

The mask b_1 is a binary sequence of length k . The adaptive encoding requires maintaining two running counters that count incrementally the number of 0s and number of 1s, starting from an initial value of 1. Traversing the bit mask left to right, for every symbol in b_1 , Alice estimates its probability by dividing the current state of the symbol's counter by the sum of the two counters. After the probability is estimated, Alice increments the corresponding counter by 1. The code length to state each symbol is the negative logarithm of its estimated probability. Generalizing this, if

$cnt[0]$ is the number of 0s and $cnt[1]$ be the number of 1s in any bit mask of size k , then the length of the message to transmit this bit mask is $-\log_2 \left(\frac{cnt[0]! \times cnt[1]!}{(k+1)!} \right)$ bits. Figure 1(a) gives an example. Notice that both Alice and Bob initialize their counters to 1. Alice encodes each symbol in the bit mask and transmits it before incrementing the corresponding counter at her end. Bob decodes the received symbol using the same estimate of the probability and updates the counters on his side, thus keeping both counters synchronized to achieve a lossless communication.

Step 4: Transmitting absolute positions in τ_2 of the overlapping elements.: This again defines another bit mask, b_2 . It is easy to see that there are $\binom{k}{cnt[1]}$ possible candidates for b_2 , given that Bob already knows b_1 . Therefore, assuming these candidates are uniformly distributed, the optimal message length to state b_2 takes $\log \left(\binom{k}{cnt[1]} \right)$ bits. We emphasize here that b_2 ignores the permutation of the overlapping elements as they appear in τ_2 (with respect to τ_1) – this is handled in the next step. While the above encoding is optimal, it, however, does not account for the displacement of overlapping elements in terms of their absolute ranks in the list. It might arise in some applications that the displacement is among the criteria of comparing two lists. Hence we propose a modified adaptive scheme to account for this displacement. We use two counters; the first tracks the number of times the symbols in bit masks b_1 and b_2 remain the same at a given position (column); the second tracks the number of times they are different. These counters are used to estimate the probabilities while traversing along b_2 . For example, See Figure 1(b).

Step 5: Transmitting the permutation of overlapping elements in τ_2 with respect to τ_2 : From the previous step, Bob knows what the overlapping elements between the lists are, but does not know in what order they appear in τ_2 . To transmit the permutation of these overlapping elements efficiently, a lexicographic numbering can be mutually agreed between them (as a part of the code book). Then, transmitting the permutation of these elements requires simply communicating its lexicographic number over some integer code.

However, to make this transmission efficient a factoradic (or mixed factorial base numbering system) can be employed [19]. This system defines a bijection between the symmetric group S_n to $n!$ possible permutations in that group.

Concretely, let $\pi = \{\pi(1), \pi(2), \dots, \pi(n)\}$ be some permutation of n symbols, where $\pi(i)$ is the rank of the i th element in the permutation. A factoradic of π defines a sequence $f(\pi) = (f^1, f^2, \dots, f^n)_!$, where any f^i is the number of j s greater than i such that $\pi(i) < \pi(j)$. See Figure 2 for an example of a lexicographic ordering of the symmetric group S_4 labeled by elements 'a, b, c, d', along with its corresponding factoradic sequence of digits.

It can be observed that each factoradic digit f^i denotes the number of successive *adjacent transpositions* on π required to move each $\pi(i)$ th element into its correct position. The permutation index (in decimal) can be computed from a factoradic as $\sum_{i=0}^n f^i \times (n-i-1)!$.

The sequence of digits $f(\pi)$ has several interesting properties. It has been shown that, if permutations in S_n are distributed uniformly, each factoradic digit f^i is also uniformly distributed in the range $0 \leq f^i \leq (n-i-1)$ [20]. Also, the factoradic digits f^i are mutually independent of each other because they form projections on independent factors in the product $n \times (n-1) \times \dots \times 1 \equiv n!$.

Thus, transmitting a permutation of overlapping elements in τ_2 involves transmitting its factoradic digits in sequence. For each factoradic digit f^i in the range $0 \leq i < n$ (note: f^n is always 0), any decreasing probability distribution on integers in that range can be used. Specifically, we use a Wallace tree code [21] that defines a code over positive integers³ by associating each integer with the binary code used to uniquely identify a binary tree. Since this integer code is defined over the infinite space of positive integers, the probability associated with each code is normalized such that the total probability in the finite range $0 \leq f^i \leq n-i-1$ adds up to 1.

Step 6: Transmitting non-overlapping elements in τ_2 . Given that the domain of elements is known and is of size N , each non-overlapping element can be stated in $\log_2(N - |\tau_1 \cup \tau_2|)$ bits. With this the communication process concludes.

A. Case 2: When the domain of elements is unknown

Here Alice and Bob do not know the domain of elements being sorted. In the previous case, Steps 1,2 and 6 depended on knowing the domain, and hence require modification. The encodings for Steps 3,4, and 5 remain exactly the same as previously described.

Since this framework relies on lossless transmission, and there is no prior knowledge of the domain of possible labels in each of the two top k lists, this requires the lists (along with its labels) to be explicitly communicated.

To efficiently communicate τ_1 and the non-overlapping elements in τ_2 , consider the union of the two lists, $\tau_1 \cup \tau_2$, such that the top k elements define labels in τ_1 (in that order) and the remainder are the labels of non-overlapping elements in the order they appear in τ_2 .

First, the size of the union $|\tau_1 \cup \tau_2|$ is transmitted using the Wallace tree code defined over all positive integers. (This modifies previous Step 1.) Then the labels in the $\tau_1 \cup \tau_2$ can be compressed using, for instance, a standard, dictionary-based lossless data compression algorithm of Lempel-Ziv-Welch (LZW) [22]. The length, $|LZW(\tau_1 \cup \tau_2)|$, in bits

³Since factoradic digits start from 0, we just add 1 to each digit to map it to the Wallace tree code.

gives the cost to state the information in τ_1 and non-overlapping elements in τ_2 . (This modifies previous Steps 2 and 6).

Computational complexity of computing various code length terms. For case 1: Code lengths involved in steps 1, 2, and 6 take $O(1)$ time to compute. The adaptive code lengths in Steps 3 and 4 take $O(k)$ time. In Step 5, finding the code length corresponding to the factoradic of a permutation of n overlapping elements can be achieved in $O(n \leq k)$ time [23]. Computing the code length of each factoradic takes $O(1)$ time. Thus, the total computational complexity to estimate the information content in the two lists grows as $O(k)$. For case 2: Step 1 requires $O(1)$ time. Steps 2 and 6 are dealt together and involves compression of labels in the set $\{\tau_1 \cup \tau_2\}$. It can easily be seen that $k \leq |\tau_1 \cup \tau_2| \leq 2k$. LZW compression implemented naively has a time complexity of $O(S|\aleph|)$, where S is the number of input symbols and \aleph is the alphabet. For most practical applications, both S and $|\aleph|$ are $O(k)$ in size. Steps 3, 4, and 5 are the same as in case 1. Thus, the total time complexity to estimate of the information content for this case grows as $O(k^2)$.

IV. RESULTS

We first quantify the effect of disarray between permutations as assessed by various popular measures. Figure 3 gives the cost associated with various measures for the set of all permutations in symmetric groups (a) S_7 and (b) S_8 . Specifically, the measures used are: (1) Spearman's foot rule metric (L_1 distance), (2) Canberra distance (weighted L_1 measure), (3) Kendall's tau distance, measuring the number of adjacent transpositions to sort a permutation, and (4) the information measure we developed in this work. It can be seen that as the information content to describe a permutation increases, all the other measures vary significantly. It is important to note that the costs reported by all four of the considered measures are related to the total number of adjacent transpositions of elements required to sort the permutation. However, our measure of information accounts for the varying magnitude of disarray (given by the permutation's factoradic digits) of each element, instead of combining and summarizing using a simple number. Other measures overlook these individual contributions; for instance, it can be seen from Figure 2 that the permutations $adcb = 5_{10} = (0, 2, 1, 0)_1$, $bcda = 9_{10} = (1, 1, 1, 0)_1$, $bdca = 10_{10} = (1, 2, 0, 0)_1$, $cadb = 13_{10} = (2, 0, 1, 0)_1$, and $dabc = 18_{10} = (3, 0, 0, 0)_1$ all require the same number of transpositions ($= 3$), yet differ in the number of transpositions for its individual elements.

To examine the performance of various measures on comparing top k lists, we first consider three top 250 movie lists downloaded from goodmovieslist.com, imdb.com and reddit.com. Figure 4 shows the comparisons of (left to right) Goodmovies vs. IMDb, Goodmovies vs. Reddit, and IMDb

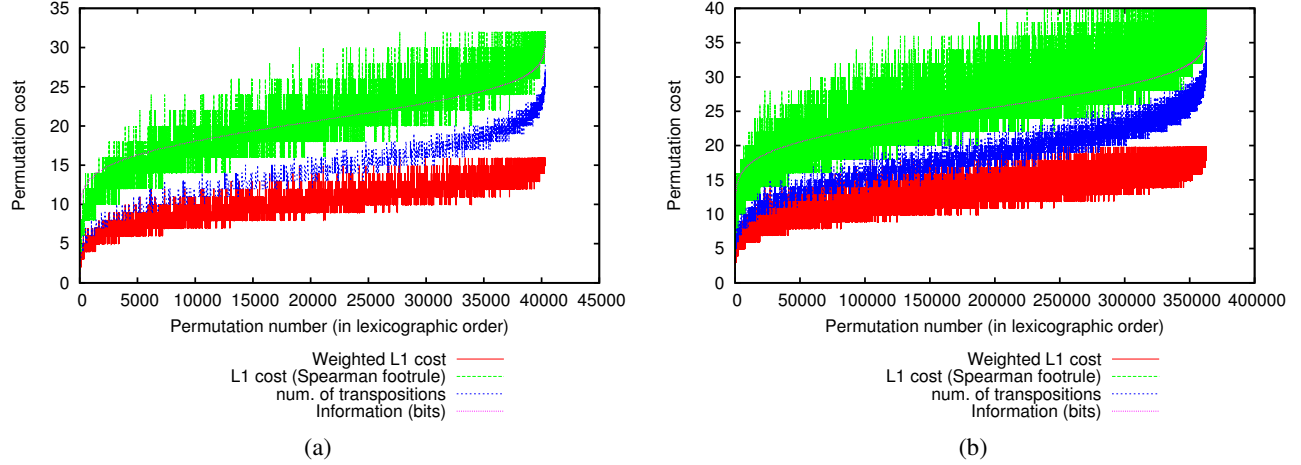
b_1	0 0 1 1 0 0 1 0 0 0	b_1	0 0 1 1 0 0 1 0 0 0
$cnt[0]$	1 2 3 3 3 4 5 5 6 7	b_2	0 0 1 1 1 0 0 0 0 0
$cnt[1]$	1 1 1 2 3 3 3 4 4 4	$cnt[0 0]$ or $cnt[1 1]$	1 2 3 4 5 5 6 6 7 8
Prob.	$\frac{1}{2} \frac{2}{3} \frac{1}{4} \frac{2}{5} \frac{3}{6} \frac{4}{7} \frac{3}{8} \frac{5}{9} \frac{6}{10} \frac{7}{11}$	$cnt[0 1]$ or $cnt[1 0]$	1 1 1 1 1 2 2 3 3 3
		Prob.	$\frac{1}{2} \frac{2}{3} \frac{3}{4} \frac{4}{5} \frac{1}{6} \frac{5}{7} \frac{2}{8} \frac{6}{9} \frac{7}{10} \frac{8}{11}$

(a)

(b)

Figure 1. Examples of the adaptive encoding schemes for bit masks described in the main text.

0_{10}	a b c d (0 0 0 0) _i	6_{10}	b a c d (1 0 0 0) _i	12_{10}	c a b d (2 0 0 0) _i	18_{10}	d a b c (3 0 0 0) _i
1_{10}	a b d c (0 0 1 0) _i	7_{10}	b a d c (1 0 1 0) _i	13_{10}	c a d b (2 0 1 0) _i	19_{10}	d a c b (3 0 1 0) _i
2_{10}	a c b d (0 1 0 0) _i	8_{10}	b c a d (1 1 0 0) _i	14_{10}	c b a d (2 1 0 0) _i	20_{10}	d b a c (3 1 0 0) _i
3_{10}	a c d b (0 1 1 0) _i	9_{10}	b c d a (1 1 1 0) _i	15_{10}	c b d a (2 1 1 0) _i	21_{10}	d b c a (3 1 1 0) _i
4_{10}	a d b c (0 2 0 0) _i	10_{10}	b d a c (1 2 0 0) _i	16_{10}	c d a b (2 2 0 0) _i	22_{10}	d c a b (3 2 0 0) _i
5_{10}	a d c b (0 2 1 0) _i	11_{10}	b d c a (1 2 1 0) _i	17_{10}	c d b a (2 2 1 0) _i	23_{10}	d c b a (3 2 1 0) _i

Figure 2. All possible permutation of elements a, b, c, d , their lexicographical number in base 10, and their corresponding sequence of digits in a factorial number system. The factoradic system defines a bijection between the permutation and its lexicographic number. For example, $21_{10} = (3, 1, 1, 0)_i = 3 \times 3! + 1 \times 2! + 1 \times 1! + 0$.Figure 3. Variation of costs over the set of all permutations in the symmetric groups (a) S_8 , and (b) S_9 . Spearman's foot rule distance is in Green. Kendall tau distance is in Blue. Canberra distance is given in Red. Information measure defined in this work is given in Magenta.

vs. Reddit, varying k from 1 to 250 in increments of one. Qualitatively the lists corresponding to Goodmovies and IMDb are more similar than the other possible pairs. It can be seen from the figure that both Spearman's foot rule distance and Kendall tau distance grow roughly quadratically with the size of k . This mainly results from the contributions to the respective costs from the set of non-overlapping elements. As this set grows, its contribution to the distance

dominates. However, the growth of information distance⁴ is roughly linear. This makes more sense, as information is additive. When the size of the list increases from k to $k+1$, the new element that gets added to each of the two lists can in the worst case be independent of the previous information. This implies that, in the worst case, the total information content in the list going from k to $k+1$ gets augmented by

⁴We note that, for these results and those to follow, we compute the measure of information between lists by assuming that the total domain of movies is unknown, i.e., following case 2 described in Section III-A.

the sum of information in the new elements. Therefore, the quadratic growth of the other measures is questionable.

It is interesting to note that while Spearman's foot rule and Kendall tau distances monotonically increase, the information distance plotted in the figure has 'fluctuations' in the amount of information measured. These variations occur when new elements (for increasing values of k) cause the set of overlapping elements to grow in size. While this increases the distance to state the permutation of overlapping elements, there is a net saving because the size of the set of non-overlapping elements (which are transmitted using *LZW* compression) decreases, in comparison with the previous values of k .

Further, to undertake this comparison on a larger scale, we compare the search results of three popular web search engines: Google, Yahoo and Ask. We achieve this by selecting 250 top trending search and news terms reported by Google Trends and Yahoo text Analytics for the regions of Australia, US, India, Canada, UK, Singapore and Germany. Figure 5 plots the average (mean) distance over all the 250 queries computing using Information, Spearman's foot rule and Kendall tau measures. In this experiment, we vary k more coarsely as 10, 25, 50, 75, and 100. In this figure the same growth trends from before emerge, linear growth for information distance and quadratic growth for Spearman's and Kendall's distance. For $k > 50$, the difference between the average distances grows drastically for Spearman's and Kendall's distances, while the same using information distance does not. This suggests a major shortcoming of these measures compared to our measure based on information.

V. CONCLUSION

We have introduced a new information measure for comparing any two top k lists. By exploring their compressibility, our method provides a statistically rigorous measure of variability between ranked lists. It provides an objective trade-off between criteria that measure the dissimilarity between lists, addressing the lacunae and pitfalls in the existing measures. As a future direction of research, this measure can be used to address the important *rank aggregation problem*: What is the 'consensus' top k ranking that combines the top k results from multiple sources.

VI. ACKNOWLEDGMENTS

We thank Chetana Gavankar for rekindling our interest on this problem. We thank Lloyd Allison for numerous discussions and helpful suggestions on this topic.

JHC's doctoral research is supported by Australian Postgraduate Award (APA) and NICTA PhD scholarship. NICTA is funded by the Australian Government through the Department of Communications and the Australian Research Council through the ICT Centre of Excellence Program.

REFERENCES

- [1] R. Fagin, R. Kumar, and D. Sivakumar, "Comparing top k lists," *SIAM Journal on Discrete Mathematics*, vol. 17, no. 1, pp. 134–160, 2003.
- [2] C. Spearman, "The proof and measurement of association between two things," *The American journal of psychology*, vol. 15, no. 1, pp. 72–101, 1904.
- [3] M. Kendall, "A new measure of rank correlation," *Biometrika*, vol. 30, no. 1, pp. 81–93, 1938.
- [4] R. Fagin, R. Kumar, M. Mahdian, D. Sivakumar, and E. Vee, "Comparing partial rankings," *SIAM Journal on Discrete Mathematics*, vol. 20, no. 3, pp. 628–648, 2006.
- [5] J. Bar-Ilan, M. Mat-Hassan, and M. Levene, "Methods for comparing rankings of search engine results," *Computer Networks*, vol. 50, no. 10, pp. 1448–1463, 2006.
- [6] E. Budinska, K. Kugler, and S. Lin, "Package topklists for rank-based genomic data integration," in *Proceedings of IASTED Computational Biology*, M. G. Schimek, Ed., 2011.
- [7] W. Fury, F. Batliwalla, P. K. Gregersen, and W. Li, "Overlapping probabilities of top ranking gene lists, hypergeometric distribution, and stringency of gene selection criterion," in *IEEE Conference on Engineering in Medicine and Biology Society*, 2006, pp. 5531–5534.
- [8] R. Pearson, "Reciprocal rank-based comparison of ordered gene lists," in *IEEE Workshop on Genomic Signal Processing and Statistics workshop*, 2007, pp. 1–3.
- [9] G. Jurman, S. Riccadonna, R. Visintainer, and C. Furlanello, "Canberra distance on ranked lists," in *Proceedings, Advances in Ranking-NIPS 09 Workshop*, 2009, pp. 22–27.
- [10] —, "Algebraic comparison of partial lists in bioinformatics," *PloS One*, vol. 7, no. 5, p. e36540, 2012.
- [11] G. Lance and W. Williams, "Computer programs for hierarchical polythetic classification (similarity analyses)," *The Computer Journal*, vol. 9, no. 1, pp. 60–64, 1966.
- [12] C. S. Wallace and D. M. Boulton, "An information measure for classification," *Computer Journal*, vol. 11, no. 2, pp. 185–194, 1968.
- [13] C. S. Wallace, *Statistical and Inductive Inference using Minimum Message Length*, ser. Information Science and Statistics. SpringerVerlag, 2005.
- [14] A. N. Kolmogorov, "On tables of random numbers," *Sankhyā: The Indian Journal of Statistics, Series A*, vol. 25, no. 4, pp. 369–376, 1963.
- [15] R. J. Solomonoff, "A formal theory of inductive inference. part i," *Information and control*, vol. 7, no. 1, pp. 1–22, 1964.
- [16] G. J. Chaitin, "On the length of programs for computing finite binary sequences," *Journal of the ACM (JACM)*, vol. 13, no. 4, pp. 547–569, 1966.

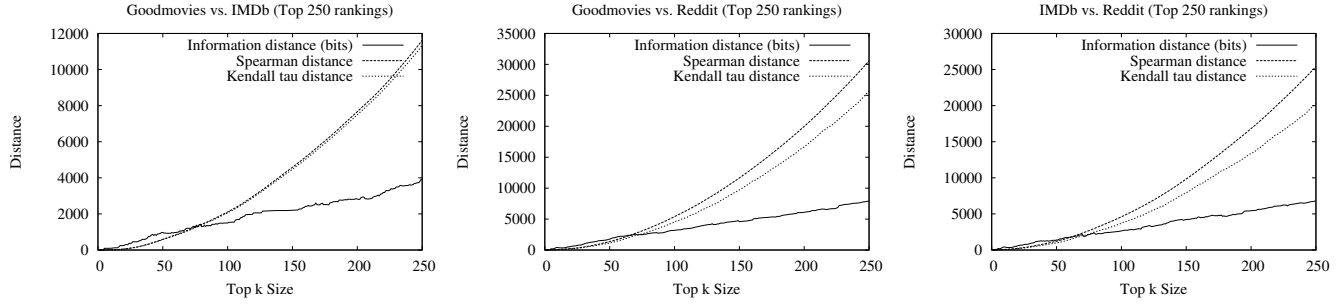


Figure 4. Comparison of Information distance, Spearman's distance, Kendall's tau distance on movie rankings from `goodmovieslist.com`, `imdb.com` and `reddit.com`

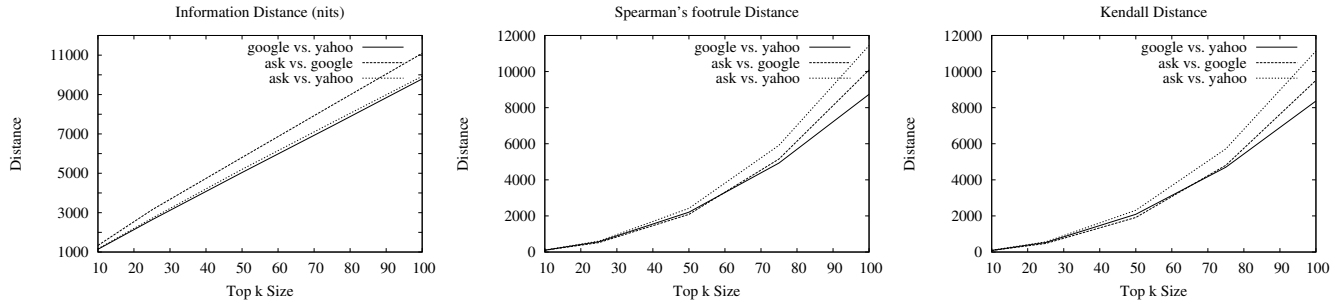


Figure 5. Comparison of Information distance, Spearman's distance, and Kendall's tau distance on search results of Google, Yahoo and Ask. The reported values are averaged over 250 top-trending search terms comparing pairs of ranked lists for values of $k = \{10, 25, 75, 100\}$

- [17] C. E. Shannon, "A mathematical theory of communication," *Bell System Technical Journal*, vol. 27, pp. 379–423, 1948.
- [18] T. Bayes and R. Price, "An essay towards solving a problem in the doctrine of chance," *Philosophical Transactions of the Royal Society of London*, vol. 53, pp. 370–418, 1763.
- [19] D. Knuth, *The art of computer programming*, 1999, vol. 3.
- [20] D. Lehmer, "Teaching combinatorial tricks to a computer," in *Proc. Sympos. Appl. Math. Combinatorial Analysis*, vol. 10, 1960, pp. 179–193.
- [21] C. Wallace and J. Patrick, "Coding decision trees," *Machine Learning*, vol. 11, no. 1, pp. 7–22, 1993.
- [22] J. Ziv and A. Lempel, "Compression of individual sequences via variable-rate coding," *Information Theory, IEEE Transactions on*, vol. 24, no. 5, pp. 530–536, 1978.
- [23] W. Myrvold and F. Ruskey, "Ranking and unranking permutations in linear time," *Information Processing Letters*, vol. 79, no. 6, pp. 281–284, 2001.