BENCHMARKING WITHOUT GROUND TRUTH

Simone Santini

Escuela Politécnica Superior, Universidad Autónoma de Madrid
Madrid, Spain

ABSTRACT

Many evaluation techniques for content based image retrieval are based on the availability of a ground truth, that is on a "correct" categorization of images so that, say, if the query image is of category A, only the returned images in category A will be considered as "hits." Based on such a ground truth, standard information retrieval measures such as precision and recall and given and used to evaluate and compare retrieval algorithms. Coherently, the assemblers of benchmarking data bases go to a certain length to have their images categorized.

The assumption of the existence of a ground truth is, in many respect, naïve. It is well known that the categorization of the images depends on the a priori (from the point of view of such categorization) subdivision of the semantic field in which the images are placed (a trivial observation: a plant subdivision for a botanist is very different from that for a layperson). Even within a given semantic field, however, categorization by human subjects is subject to uncertainty, and it makes little statistical sense to consider the categorization given by one person as the unassailable ground truth.

In this paper I propose two evaluation techniques that apply to the case in which the ground truth is subject to uncertainty. In this case, obviously, measures such as precision and recall as well will be subject to uncertainty. The paper will explore the relation between the uncertainty in the ground truth and that in the most commonly used evaluation measures, so that the measurements done on a given system can preserve statistical significance.

1. INTRODUCTION

As content based image retrieval enters its second decade, the problem of a scientifically correct and repeatable evaluation methodology keeps pressing with the urgence of the important yet unsolved problems. The general difficulty of evaluating content based image retrieval systems derives from a number of problems, only some of which are technical (issues such as the increasing reach of copyright laws, the dwindling of "fair use," and the consequent problematicity in asembling large, publicly available, test data bases comes to mind as a very relevant non technical problem).

In this article, I am concerned with one rather specific methodological point of image retrieval benchmarking, namely, the reference to a *ground truth* against which the performance of a retrieval system is measured. In many cases, the evaluation of image retrieval systems is based on measures derived from information retrieval, such as precision/retrieval curves, which assume the existence of an *a priori* categorization, valid absolutely, that is, valid independently of the measurement operation, of the query process, and on the person whi is either doing the query or the categorization. (I will call this a *strong* ground truth.)

It is my intention to argue the problematicity of such an absolutist notion of ground truth. In particular, it is my intention to argue against the notion of a taxonomy that can be, in principle, derived from the image data and that would consequently be independent of the person who creates it or of the circumstances under which it is created.

If my argument is valid, then we need to revise our evaluation criteria, at least those based on a *strong* ground truth. In this paper I will do this by moving along two directions: firstly, by defining some measurement methodologies that do not depend on a "strong" notion of ground truth, secondly, by defining variants of the standard measures that can deal with uncertainty in the ground truth.

The remainder of this paper is organized as follows. Section 2 is a brief introduction to the problems of benchmarking, to the different types of evaluation that are available to the experimenter, and to the problems associated to their use. Section 3 will review the most common measures based on ground truth used in evaluation,

Internet Imaging VII, edited by Simone Santini, Raimondo Schettini, Theo Gevers,
Proc. of SPIE-IS&T Electronic Imaging, SPIE Vol. 6061, 60610I, © 2006 SPIE-IS&T · 0277-786X/06/$15

and discuss the problems of principle that problematize their application. Secions 4 and 5 will present the two techniques proposed here to overcome the problems posed by the definition of ground truth. Section 6 will prsent some conclusions.

2. BENCHMARKING WHAT?

There are, of course, many axes along which one might want to measure the work done by a content based retrieval system, and not all of them are of interest in this paper. An important class of measurement of a system hs to do with what we might call its *computational performance,* that is, with the measure of the efficiency with which the system uses the two fundamental resources of time and memory space. These measures are completely objective and the measurement units in which they are expressed (seconds and bytes) can be determined completely independently of the presence of an observer. There is, in other words, no observer-given "ground truth" to speak of. It must be noted, however, that even in this case, the full objectivity of the ultimate, so to speak, evaluation, is questionable. A computer system is not created *in vacuo,* but it is designed to perform a certain function in an organization.

A true perfoarmance evaluation, then, should take into account not much (or, at least, not only) the *physical* performance of the system, but its efficiency in terms of the organization in which it is inserted. The effect of the introduction of a system must be compared *vis á vis* the way in which the same function was performed, within a particular organization, before the system was introduced. So, one might want to compare the performance of a fully automated information system that provides information to the customer by a telephone menu against the presence of a customer representative that provides the same information. The problem, however, is that the very presence of the computer system changes the organization so that, while the specific operation might become more efficient, the whole organization might not, or vice-versa. In the case of the computer system for answering customers, the simplest way of evaluating the performance of the system would be to measure the time necessary to retrieve a given piece of information. In this sense, the system certainly outperforms a person doing the same job.

Once the system is inserted into the customer service organization, however, the organization changes: it typically becomes more rigid, in such a way that, while the answer to queries foreseen by the designers of the system is very rapid, taking care of an unforeseen or complicated customer problem can be frustrating or nearly impossible. Measuring the time necessary to retrieve an information item is in this case misleading, since a customer almost never needs to retrieve a specific information item: the customer has a problem and, in general, lacks the knowledge of the system necessary to describe it precisely. The general knowledge of the organization that a customer representative has, and her flexibility in filling in the missing information from the context of the customer interaction, might be in this case more important than the speed at which information is retrieved. It is a fact that, in general, the creation of an automatic telephone system for customer interaction results in a growing customer dissatisfaction with the service.

In this case, a seemingly objective measure (retrieval time) predicted an improvement but, once the system is inserted in the organization in which it has to work, the structure of the organization change in the direction of a lower efficiency. The morale is, of course, that one should be wary of completely objective measures: while they are useful in order to compre systems with each other (*coeteris paribus* it is certainly better to install a fast system than a slow one), the ultimate effect of a system on an organization is more complex to define and measure.

A class of measures more directly related to the place that a computerized system will occupy in an organization is given by what we might call *functional measures,* that is, measures that evaluate the functional behavior of the system, its adherence to a given functional specification. In the case of a visual information retrieval system, for instance, the purpose of a system can be to retrieve images following certain criteria, criteria that are usually specified in cognitive or semiotic terms. So, apart from all the important technical measures such as the time necessary to retrieve an image in a data bse of a given size, or the size of the index that allows fast retrieval, apart from all this, the essential question that benchmarking is supposed to answer is: how well this system perform against other systems in the fulfillment of its function?

The same arguments advanced before will lead us to acknowledge the contextuality of such an evaluation. In a computerized system (or in a system of any other kind) that is inserted in an organization, there is no such a thing as an a-contextual function. In this sense, in the area of visual information retrieval it is impossible to evaluate the system *vis á vis* its formal requirements and then, independently, its formal requirement *vis á vis* the organization. The reason for this state of affairs is that the requirements of the system have always a cognitive component and, as such, are never fully attainable.

An example will serve to clarify this point. Consider a visual information system installed in a legal firm dealing with trademarks and intellectual property. There are many components of such a system but let us focus on the feature extractor and similarity measure that determines whether two trademarks are too similar to be acceptable or not. The system, as a whole, will have a number of technical requirements, having to do with response speed, storage capacity, security of the data, etc. The search subsystem, on the other hand, is expected to perform as well as a lawyer engaged in the analysis of trademarks. In other words, its requirements are expressed as a cognitive capacity.

This brief excursus in performance measure served well, I believe, as a warning against a simplistic attitude in benchmarking that tends to highlight the technical problems and relegate the system context to a secondary place. It is important to acknowledge the importance of the specific application in which a system is employed because this acknowledgment makes evident the tension that exists in benchmarking: while evaluation of a complex information system is always contextualized by the organization in which it is placed, and while the evaluation of functions of a system are always contextualized by the whole in which they are placed, while all this is true, a benchmark must try to achieve a certain generality, since it is to be used for the comparison of disparate systems.

In this tension, an important rôle is played by the ground truth: as it is defined and used in many benchmarking data bases, the ground truth is an absolute *datum,* that is, it is completely de-contectualized. Any way to alleviate this absoluteness of the ground truth will therefore work in the direction of conciliating the contrasting requirements of generality and contectual evaluation.

3. MEASURES BASED ON A STRONG GROUND TRUTH

The most common measures used in content based image retrieval are rather direct translations of the homologous measures used in information retrieval. This fact is, *prima facie,* not surprising, since information retrieval is, in spirit if not in techniques, quite close to content based image retrieval, and bears a very direct influence on it. Information retrieval is concerned with searches is free text (unstructured) documents and, as such, it has to deal with the problem of inferring semantics from its syntactic "traces," which is, *mutatis mutandis,* the same problem that content based image retrieval has to contend with. The most common measures that have passed from information retrieval to content based image retrieval are the so-called *recall* and *precision*[*]. Given a data base D and a query, let $V \subseteq D$ be the subset of documents that are *relevant* to the query. Suppose now that a given system returns, for the same query, the set of answers $A \subseteq D$. The *recall* is the fraction of the relevant documents that the system returns:

$$\mathcal{R} = \frac{|A \cup V|}{|V|} \in [0,1] \tag{1}$$

A value $R = 1$ means that all relevant documents have been retrieved by the system, a value $R = 0$ means that no relevant document was retrieved. (The recall is undefined if the data base contains no relevant documents.) This measure alone is quite clearly insufficient to characterize the quality of a retrieval system; suffice it to say that any system that returns the whole data base (that is, any system for which $A = D$) would have $R = 1$. The *precision* measures the fraction of the returned documents that are relevant:

$$\mathcal{P} = \frac{|A \cup R|}{|A|} \in [0,1] \tag{2}$$

[*] Equivalent measures have been used in a number of fields, with slightly different definitions and wildly different denominations. For example, in medical research, measures based on almost the same definition are called *sensitivity* and *specificity.* The names precision and recall, however, appear to be the most common in content based image retrieval, and I will use them throughout the paper.

Usually, increasing the size of A (i.e. returning more elements) increases recall and decreases precision.

The epistemologically problematic assumption here is the possibility of defining the set V, that is, the possibility of dividing neatly the data base in two parts, the first of which is completely relevant to the query, the second of which is completely irrelevant. The applicability of this hypothesis to general queries entails two crucial hypotheses:

i) the data base can be categorized in a way that all subjects will find unambiguous;

ii) the queries are about categories that, if they do not completely coincide with those of the categorization, are at least cut along the same semantic axes.

Benchmarking has adapted, by and large, to these theoretical presuppositions by providing, on one hand, categorized data bases and, on the other hand, by encouraging (or, at least, accepting) experimental methodologies based on the categories in which the data base was divided. To make but an example, one of the most commonly used data bases in content based image retrieval, the Corel data base, is structured as a taxonomy, and in the virtual totality of the article that make use of this data base in experiments, the "test query" corresponds to one of the categories of the data base.

But there are serious reasons to doubt of both these hypotheses. Categorization is a form of interpretation, that is, a form of reading and, as literary theory teaches us, the result of reading is as much a function of a reader than it is of the text that is being read. To name only one aspect, "readers do not of course encounter texts in a void: all readers are socially and historically positioned, and how they interpret literary works will be deeply shaped by this fact"[2]; reading is not just a matter of receiving the text, as much as of changing it, interpreting it: the true writer is the reader.[3] The same considerations are all the more valid for images and, especially, for that specific reading act that consists in categorizing the images in a data base. In this case, one would argue, we are in the presence of the visual equivalent of the ultimate post-modern text, a *pastiche* of fragments without a thread that might somehow guide and constraint the reader's freedom to interpret.

To expect, in this situation, that different readers will divide the data base (the text) in the same categories would be to expect too much!

We must also contend with the fact that, in a realistic situation, a query will not necessarily be asked along the same semantic axes used to derive the categorization. As a simple example, many benchmarks start their taxonomies with a division of images in indoor and outdoor scenes and then, later on, switch the semantic axis from the setting to, say, the objects contained in the image, dividing the sub-sections obtained so far in images with animals, with persons, and so on.

But a query might go along completely different semantic axes: a query might ask for images in which the sky is cloudy[†], for images that remind one of Mr. Pickwick, or for any other thing.

The presence of a categorization, in other words, limits the freedom of the experimenter to choose the queries with which the evaluation will be done. This, in turn, can bias the evaluation of a system for, clearly, a system designed from the beginning with in mind the same semantic categorization axes as the benchmark data base will perform better than a system designed for a completely unrelated division of the semantic field.

One of the problems here is that the categorization is *a priori,* so that the system, whose query takes place in a given context, is not compared with a single subject in the same context, but with a de-contextualized

[†] Note that this does not mean that the query is aligned with the exterior/interior axis of the classification, since one might very well be interested in images taken indoors where the cloudy sky is seen from a window.

assembly of subjects. The strategy can be illustrated in the following diagram:

$$\begin{tikzcd}
u_1 \arrow[rdd] & & \\
u_2 \arrow[rd] & & \\
\vdots & B \arrow[rdd] & \\
u_n \arrow[ru] & & \\
& & M \\
Q \arrow[r] & X \arrow[ru] &
\end{tikzcd}$$

(3)

Here a group of subjects $u_1, \ldots, u_n$ are used to obtain the categorized benchmark B, independently of the system evaluation. Then, an evaluation trial consists in using a query Q on the system under evaluation X, and comparing the results with the categorization B. The problem, as we have seen, is that in this schema, the *context* in which each subject contributed to the categorization, that is, the specific reading that each user gave of the data base is lost and plays no rôle in the determination of M: the output of X is not compared with the situated answer of each subject u_i, but with an agglomerate B whose epistemological validity I have just questioned.

There is no absolute way to escape to this situation: the readings of the subjects are contextual, and we can't formalize the cultural context to make it available to the system X. There are, however, two directions that we can take to, at least, attenuate the effects of de-contextualization. The first, that I will discuss in the next section, is to relax the absolutist assumptions on the categorization B, that is, to encode, if not the context in which the different evaluations are made by the u_i, at least the fact that there will be discrepancies in the assignment of images to categories, and that the greatest the discrepancy in the assignment to a category, the less one should rely on the semantic value of that category.

The second direction, which I will discuss in the following section, is to rely on a different type of evaluation schema, one such as the following:

$$\begin{tikzcd}
& u_1 \arrow[r] & R_1 \arrow[r] & M_1 \arrow[rd] & \\
Q \arrow[ru] \arrow[rd] \arrow[rdd] & \vdots & \vdots & \vdots & M \\
& u_n \arrow[r] & R_n \arrow[r] & M_n \arrow[ru] & \\
& X \arrow[r] & R \arrow[ruuu] \arrow[ru] & &
\end{tikzcd}$$

(4)

Here the *same* query is posed to the subjects and to the system, and a single trial consists in comparing the answer of the system with that of each subject to obtain individual discrepancy measures, which are then combined in a single discrepancy measure. Note that in this case the comparison is made with each single subject, that is, in the context created by the query Q. The aggregation of the results across the different subjects only takes place after they have been converted to a numeric value.

Note that, from the point of view of the builder of a benchmarking data base, this solution has the considerable disadvantage of being much more expensive to create, and of requiring the storage of many more data *vis á vis* a categorized data base. In this case, in fact, it will be necessary for the data base to store not only images, but a collection of queries as well and, for each one of these queries, the response of all the subjects. As a final aside, I will notice that this solution has at least the advantage of making explicit the dependence of evaluation on the specific queries that are permitted. As mentioned before, the act of categorizing along certain semantic axes, in itself, entails a restriction in the form of the queries that can be used for evaluation and, therefore, a bias towards certain types of systems. Categorization alone, however, lulls the experimenter in the false illusion of being the arbiter of the queries that are used in the experiment. Making the queries part of the data base will, if nothing else, make this dependence explicit.

The greatest cost of this solution, however, plays in favor of keeping, if not the categories as rigidly defined as they are today, at least the experimental schema of categorization, which is the solution I will discuss in the next section.

4. UNCERTAINTY IN THE GROUND TRUTH

If we want to maintain the theoretical structure of comparison with a prescribed categorization, as illustrated in (3) we must, at least, acknowledge the problematicity of the categorization by making it less crisply defined. One way to do this is to give it a probabilistic interpretation, that is, each image will potentially belong to all the prescribed categories, belonging to each one with a certain probability, given by the fraction of subject that assigned the given image to the given category.

I will assume that the taxonomy is given in the form of a tree (the more complicated case in which the taxonomy is an acyclic graph can be tackled with techniques similar to those presented here) in which the parent/child relation ($\omega \leftarrow \nu$, where $\pm ga$ is the parent and ν is the child) means membership subsethood. In the case of standard taxonomy, this means that, for each image i and nodes (categories) ω, ν,

$$i \in \mathrm{dom}(\nu) \wedge \omega \leftarrow \nu \quad \Rightarrow \quad i \in \mathrm{dom}(\omega) \tag{5}$$

In the case of probabilistic categories, this translate in a condition that I will define in the following.

I assume that a number N of subjects is available to classify a data base of images D into one of a tree of categories ω. Each subject s assigns a category $\Omega(s, q)$ to each image $q \in D$. From these assignments it is necessary to determine the values $p(\omega|q)$, that is, the probability that the image q belongs to category ω. For the leaves of the tree, this value is given simply by the fraction of users that have assigned the image to ω, i.e.

$$p(\omega|q) = \frac{\{q|\Omega(s,q) = \omega, s = 1, \ldots, N\}}{N} \tag{6}$$

Consider now a category ν that is not a leaf of the tree, and assume that this category has children $\omega_1, \ldots, \omega_n$. The probability that an image q be of category ν is the probability that the image has either been assigned to ν by some subject, a probability given by

$$p'(\nu|q) = \frac{|\{q|\Omega(s,q) = \nu, s = 1, \ldots, N\}|}{N}, \tag{7}$$

or that it is of one of the categories $\omega_1, \ldots, \omega_n$. That is,

$$p(\nu|q) = \frac{1}{N}\left[|\{q|\Omega(s,q) = \nu, s = 1, \ldots, N\}| + \sum_{\zeta|\nu\mapsto\zeta} |\{q|\Omega(s,q) = \zeta, s = 1, \ldots, N\}| \right] \tag{8}$$

Starting with eq. (6) one computes the probability of all the leaves then, applying eq. (8) from the bottom up one computes the probability of all the other categories of the tree. Note that in general the quantity $\sum_\omega p(\omega|q)$ is greater than one, since the assignment that a subject does of an image to a category ω also extends to all the categories in the path from the root to ω that is, the same image, as assigned by one subject, belongs to more than one category. This is the probabilistic counterpart of the subsethood rule (5). On the other hand, it is easy to show that the root of the tree has always probability one.

The *size* of a category is defined as

$$|\omega| = \sum_q p(\omega|q) \tag{9}$$

Consider now a query image q, and assume for the moment that it belongs to a unique category $\hat{\omega}$. Suppose that the system under evaluation returns the images $R = \{r_1, \ldots, r_n\}$ as results. The precision of this result is

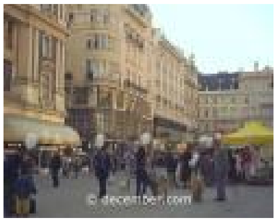


Figure 1. one of the images used for categorization

the number of images in R that belong to $\hat{\omega}$ divided by the size of R where, of course the expression "belongs to category $\hat{\omega}$" is to be intended in a probabilistic sense, and this makes precision a stochastic quantity. The probability that, out of the n elements of R, k are of category $\hat{\omega}$ is given by

$$p(k|\hat{\omega}) = \sum_{i_1<i_2<\cdots<i_k} \left[\prod_u p(\hat{\omega}|r_{i_u})) \prod_{j\neq i_1,\ldots,i_k} (1 - p(\hat{\omega}|r_j)) \right] \tag{10}$$

and the probability for the value k/n of the precision is

$$\mathcal{P}(k/n|\hat{\omega}) = \frac{p(k|\hat{\omega})}{n} \tag{11}$$

similarly for the recall:

$$\mathcal{R}(k/|\hat{\omega}|\ |\hat{\omega}) = \frac{p(k|\hat{\omega})}{|\hat{\omega}|} \tag{12}$$

All this was under the assumption that the query image q belonged to category $\hat{\omega}$ exclusively. This is in general not the case but, if the image is in the data base, we have the probabilities $p(\omega|q)$ for all categories. From these we can compute

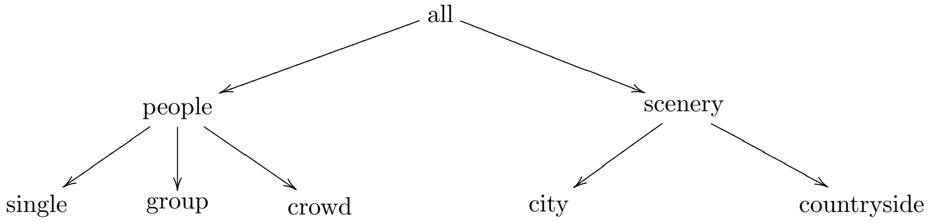
$$\mathcal{P}(k/n) = \sum_{\omega} \mathcal{P}(k/n|\omega) p(\omega|q) \tag{13}$$

and

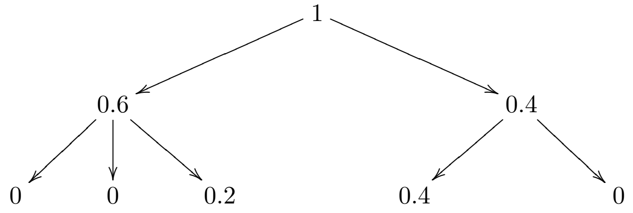
$$\mathcal{R}(k/n) = \sum_{\omega} \mathcal{R}(k/|\omega|\ |\omega) p(\omega|q) \tag{14}$$

With this definition, the precision and recall curves become distributions and the comparison of different system can be done with the standard statistical methods used to compare distributions[5,4]

Just to give an example of how the thing might work, I used a set of 20 images and asked 5 people to classify them according to the following categories (the subjects were not aware of the structure of the categories and, of course, they were shown no category named "all"):



The image of Figure 1 was classified as *people* by 2 subject, as *crowd* by one, and as *city* by two, resulting in the following probabilities for the categories:



5. INDIVIDUAL COMPARISONS

The alternative evaluation method of (4) states that the basis of evaluation should be the comparison of the answer of a subject to a specific query with the answer of the system to the same query. One advantage of this solution is that there is no need to introduce a categorization at all: we can assume that the experimental set-up is such that the answer that the subject gives is of the same type as that given by the system, and we can simply compare these two without introducing any additional infrastructure. In this kind of experiment, we consider the subject's response as the ideal, or correct, one, and we measure the data base response against it.

The details of the evaluation depend, of course, on the specific type of response that the system gives (and, consequently, on the type of response that the subjects are required to provide). If one is doing the evaluation of a system from the scratch, the type of response can be fully adapted to the characteristics of the system, but if one is building a benchmarking data base, the pre-processed subject data must be prepared with some common response in mind. Here I will consider one of the most common types of response, namely the *ranked list* of images, that is a list of images with asociated relevance measures. Let us assume for the moment that the list contains the whole data base, that is, that the answer to a query is simply an ordering of the data base. The subject response (characterized in what follows by the superscript α) will consist in a list

$$C^\alpha = [I_1, \ldots, I_n] \tag{15}$$

with associated relevances $s_i^\alpha \in [0,1]$, where $s_i^\alpha = 1$ for absolutely relevant images and $s_i^\alpha = 0$ for absolutely irrelevant images. The ordering of teh list is such that $s_i^\alpha \geq s_{i+1}^\alpha$. Since the response contains the whole data base, the system under test will respond with the same images in a different order. Using the superscript μ to characterize the response of the system, we have

$$C^\mu = [I_{\pi_1}, \ldots, I_{\pi_n}] \tag{16}$$

where $(\pi_1, \ldots, \pi_n)$ is a permutation of $(1, \ldots, n)$. The *displacement* of image k between the two orderings is given by $|k - \pi_k|$, and it is a measure of how far from the ideal rank the system under test has placed the image k. These displacements can be used to measure the quality of a system under test with a caveat: misplacing an image by a given amount doesn't always result in the same loss of quality. A system that misplaces a relevant image is considered to perform worse than a system that misplaces by the same amount an almost irrelevant image. This suggest that we might obtain an adequate measure by weighing the displacement of an image I_k by its relevance, obtaining the weighted displacement[1]

$$w = \sum_k s_k^\alpha |k - \pi_k| \tag{17}$$

If necessary, the weighted displacement can be transformed into a normalizing quality measure of the system by using a monotonically decreasing function $g : \mathbb{R}^+ \to [0,1]$. The *g-weighted quality* of the system is then

$$q = g\left(\sum_k s_k^\alpha |k - \pi_k|\right) \tag{18}$$

A class of functions for determining the normalized quality can be established as a standard using, for instance, the rational functions $g(x) = 1/(1+x)^p$ and the exponential $g(x) = \exp(-\lambda x)$.

The situation is a bit more complicated if, as itis often the case, the answer of the system (and of teh subjects) does not contain the whole data base, but only a subset that, in the case of the subject response, I will indicate as

$$C_m^\alpha = [I_1, \ldots, I_m] \tag{19}$$

Again, the images have associated a significance value, and we can assume that the significance of all the images not in C_m^α is zero. The system will also give a configuration

$$C_m^\mu = [I_{\pi_1}, \ldots, I_{\pi_m}] \tag{20}$$

but, in this case, $(\pi_1, \ldots, \pi_m)$ is not a permutation of $(1, \ldots, m)$, since some images that appear in C_m^α my fail to appear in C_m^μ, and vice-versa. The configuration can be analyzed in terms of three lists (sets ordered by the ordering induced by that of C_m^α): the list $A = C_m^\alpha \cap C_m^\mu$ of images that appear in both answers, the list $B = C_m^\alpha - C_m^\mu$ of images that appear in the answer given by the subject but not in that given by the system, and the list $C = C_m^\mu - C_m^\alpha$ of images that appear in the answer given by the system but not in that given by the subject.

The displacement is the sum of the displacements relative to the three sets. The set A can be analyzed with the same weighted displacement measure used in the previous case:

$$w^A = \sum_{I_k \in A} s_k^\alpha |k - \pi_k| \tag{21}$$

The set C is irrelevant, since its images do not belong to C_m^α and, *ex hypothesi*, their relevance is zero. The problem is the evaluation of the set B. Each image in B is a significant (for the subject) image that the system has placed beyond the range of available results. Consider an image $I_j \in B$, where j is the position in which the subject has placed it. Had the system under test returned the whole data base, the image I_j would have been placed in a position $u > m$, and its contribution to the displacement would have been $s_j^\alpha |j - u|$. Unfortately, the system only returns m images, so that there is no way of knowing the value of u.

If the system under evaluation allows incremental requests, then one can keep increasing the set C_m^μ until it includes all images selected by the subject, that is, until $C_m^\alpha \subseteq C_m^\mu$. In this case, $A = C_m^\alpha$, $B = \emptyset$, and C is irrelevant, so one can use the formula (21) to obtain the displacement.

If this is not possible, then it is necessary to make some hypothesis in order to get an approximate measure of displacement. Two reasonable hypotheses are the following:

the optimist hypothesis consists in assuming that the data base is as good as it can possibly be, compatibly with the fact that the images in B were not returned. This means that the first image in B (which, by the induced order, is the image of B with the highest rank in C_m^α), is in position $m + 1$, the second image in position $m + 2$, and so on. If β_i is the position of the ith image of B in C_m^α, then

$$w^B = \sum_{I_i \in B} s_i^\alpha |\beta_i - (m + i)|; \tag{22}$$

the pessimist hypothesis consists in assuming that, since the images in B are inaccessible, the situation is no better than it would be if the images were all placed at the end of the list, that is, the last image in B is in position N, the next-to-last in position $N - 1$, and so on. In practice, since $N \gg m$, it makes little difference where the images are placed at the bottom of the list, and one can consider a displacement N for all images in B:

$$w^B = \sum_{I_i \in B} s_i^\alpha N. \tag{23}$$

The total displacement is given by $w = w^A + w^B$, where w^B can be either the optimist or the pessimist version.

In the pessimist case note that, if the values s_k^α are of the same order of magnitude (e.g. all the images chosen by the subject are considered to be reasonably significant), then w^A/w^B is of the order of magnitude of $m/N \ll 1$, that is, $w \approx N \sum_{I_i \in B} s_i^\alpha$. In other words, in the pessimistic case, the displacement is given (apart from a factor N) by the sum of the significances of the images that the system under evaluation failed to report.

6. CONCLUSIONS

In this paper I have analyzed the rôle of a de-contextualized ground truth in the construction of benchmarks, and the problems that are associated with it. I have argued that the evaluation of an information system should always be done in the context of the organization that uses it, and that certain types of evaluation of an algorithm or a method should be done in the context of the system in which they are used.

I have presented two basic schemes for dealing with the competing needs of having a general benchmark data base against which systems can be measured and compared, and of not relying on a non contextual notion of ground truth. The first consists of modeling the context of the evaluation as uncertainty in the ground truth, which leads to the adoption of a statistical approach and to the re-definition of the standard measures of precision and recall as distributions.

The second consists in a more drastic change in the evaluation modality. The data base will store not only images, but the answer of a group of subjects to standard queries, and the evaluation will be done by comparing the output of a system with the answer of the subject in the context of a specific query, without invoking at all a general notion of ground truth.

REFERENCES

1. A. Desai Narasimhalu, M. S. Kankanhalli, and J. Wu. Benchmarking multimedia databases. *Multimedia Tools and Applications*, 4(3):333–355, 1997.
2. Terry Eagleton. *Literaty Theory; an introduction.* Minnesota University Press, 1996. (second edition).
3. Stanley Fish. *Is there a text in this class? The authority of interpretative communities.* Cambridge:MIT Press, 1980.
4. Geoffrey Keppel. *Design and Analysis. A researcher's handbook.* Prentice Hall, Upper Saddle River, NJ, 1991.
5. J. Mandel. *The statistical analysis of experimental data.* NewYork:Interscience, 1964.