# The Case of the Duplicate Documents Measurement, Search, and Science

Justin Zobel and Yaniv Bernstein

School of Computer Science & Information Technology,
RMIT University, Melbourne, Australia

**Abstract.** Many of the documents in large text collections are duplicates and versions of each other. In recent research, we developed new methods for finding such duplicates; however, as there was no directly comparable prior work, we had no measure of whether we had succeeded. Worse, the concept of "duplicate" not only proved difficult to define, but on reflection was not logically defensible. Our investigation highlighted a paradox of computer science research: objective measurement of outcomes involves a subjective choice of preferred measure; and attempts to define measures can easily founder in circular reasoning. Also, some measures are abstractions that simplify complex real-world phenomena, so success by a measure may not be meaningful outside the context of the research. These are not merely academic concerns, but are significant problems in the design of research projects. In this paper, the case of the duplicate documents is used to explore whether and when it is reasonable to claim that research is successful.

## 1 Introduction

Research in areas such as the web and information retrieval often involves identification of new problems and proposals of novel solutions to these problems. Our investigation of methods for discovery of duplicate documents was a case of this kind of research. We had noticed that sets of answers to queries on text collections developed by TREC often contained duplicates, and thus we investigated the problem of duplicate removal. We developed a new algorithm for combing for duplicates in a document collection such as a web crawl, and found that our method identified many instances of apparent duplication. While none of these documents were bytewise identical, they were often nearly so; in other cases, the differences were greater, but it was clear that the documents were in some sense copies.

However, this research outcome potentially involved circular reasoning. The existence of the problem is demonstrated by the solution, because, in large collections, manual discovery of duplicates is infeasible; and the success of the solution is indicated by the extent of the problem. That is, our algorithm succeeded on its own terms, but there was no evidence connecting this success to any external view of what a duplicate might be. We are, potentially, being misled by the use of the word "duplicate", which seems to have a simple natural interpretation. But this belies the complexity of the problem. Duplicates arise in many ways – mirroring, revision, plagiarism, and many others – and a pair of documents can be duplicates in one context but not in others.

X. Zhou et al. (Eds.): APWeb 2006, LNCS 3841, pp. 26–39, 2006.

This issue is perhaps more easily understood in an abstract case. Suppose a researcher develops an algorithm for locating documents that are *grue* (where grue is a new property of documents that the researcher has decided to investigate) and documents are defined as being grue if they are located by the algorithm. Or suppose the researcher develops an algorithm that, on some test data, scores highly for *grueness* when compared to some alternative algorithms. We can say that these algorithms succeed, but, without an argument connecting grueness to some useful property in the external world, they are of little interest.

Such problems are an instance of a widespread issue in computer science research: the paradox of measurement. We measure systems to *objectively* assess them, but the choice of measure – even for simple cases such as evaluating the efficiency of an algorithm – is a *subjective* decision. For example, information retrieval systems are classically measured by recall and precision, but this choice is purely a custom. In much research there is no explicit consideration of choice of measure, and measures are sometimes chosen so poorly that a reader cannot determine whether the methods are of value.

Thus appropriate use of measures is an essential element of research. An algorithm that is convincingly demonstrated to be efficient or effective against interesting criteria may well be adopted by other people; an algorithm that is argued for on the basis that it has high grueness will almost certainly be ignored. Problems in measurement are a common reason that research fails to have impact.

Researchers need, therefore, to find a suitable *yardstick* for measurement of the success of their solution. Yardsticks rely on assumptions that have no formal justification, so we need to identify criteria by which the value of a yardstick might be weighed. In this paper, we explore these issues in the context of our research into duplicates. We pose criteria for yardsticks and how they might be applied to duplicate detection.

Our investigation illustrates that strong, sound research not only requires new problems and novel solutions, but also requires an appropriate approach to measurement. As we noted elsewhere, "many research papers fail to earn any citations. A key reason, we believe, is that the evidence does not meet basic standards of rigor or persuasiveness" (Moffat and Zobel, 2004). Consideration of these issues – which concern the question of what distinguishes applied science from activities such as software development – can help scientists avoid some of the pitfalls encountered in research and lead to work of greater impact.

## 2 Discovery of Duplicate Documents

In 1993, not long after the TREC newswire collections were first distributed, we discovered passages of text that were copied between documents. This posed questions such as: how much plagiarism was there in the collections? How could it be found? The cost of searching for copies of a document is $O(n)$, but naïvely the cost of *discovery* of copies, with no prior knowledge of which documents are copied, is $O(n^2)$.

We developed a *sifting* method for discovery of duplicates, based on lossless identification of repeated phrases of length $p$. In this method, the data is processed $p$ times, with non-duplicated phrases of increasing length progressively eliminated in each pass: a phrase of, say, four words cannot occur twice if one of its component phrases of length

three only occurs once. In our recent refinement of this method (Bernstein and Zobel, 2004), a hash table of say one billion 2-bit slots is used to identify phrase frequency, allowing false positives but no false negatives. When all $p$-word repeating phrases have been identified, these are processed to identify pairs of documents that share at least a specified amount of text.

However, in our experiments we observed a vast number of cases of reuse of text, due to factors such as publication of the same article in different regions on different days. Cases of plagiarism – if there were any – were hidden by the great volume of other material. Moreover, the method did not scale well. In 2003 we returned to work on this problem, inspired by issues in management of large corporate document repositories, where it is common for documents such as policies and manuals to be present many times in multiple official versions, and for authors to have their own inconsistent versions. These documents represent corporate memory, yet management of them in practice may be highly chaotic; duplicate detection is a plausible method of helping to bring order to such collections. We refined our original sifting method and proposed metrics for measuring the degree of duplication between two documents.

Using the TREC `.gov` crawls, we found, disturbingly, that our metric for measuring duplication led to a smooth, undifferentiated range of scores: there was no obvious threshold that separated duplicates from non-duplicates. We had naïvely assumed that pairs would either be largely copied, with say 70% of their material in common, or largely different, with say no more than 20% in common. This assumption was entirely wrong. And again, we failed to find the kinds of duplicates we were seeking. Amongst the millions of documents there were millions of pairs (a collection of a million documents contains half a trillion potential pairs) with a reasonable amount of text in common. The diversity of kinds of duplication, rather than algorithmic issues, was the main obstacle to success. For web data, potential sources of duplicates include:

- Mirrors.
- Crawl artifacts, such as the same text with a different date or a different advertisement, available through multiple URLs.
- Versions created for different delivery mechanisms, such as HTML and PDF.
- Annotated and unannotated copies of the same document.
- Policies and procedures for the same purpose in different legislatures.
- Syndicated news articles delivered in different venues.
- "Boilerplate" text such as licence agreements or disclaimers.
- Shared context such as summaries of other material or lists of links.
- Revisions and versions.
- Reuse and republication of text (legitimate and otherwise).

At the same time as our original work, *fingerprinting* methods for duplicate detection were being developed by other groups (Manber, 1994, Brin et al., 1995, Heintze, 1996, Broder, 1997, Chowdhury et al., 2002, Fetterly et al., 2003). Several groups developed methods that broadly have the same behaviour. Some phrases are heuristically selected from each document and are hashed separately or combined to give representative keys. Two documents that share a key (or a small number of keys) are deemed to be duplicates. As most phrases are neglected, the process is lossy, but it is relatively easy to scale and is sufficient to detect pairs of documents that share most of their text.

Our sifting method can be seen as lossless but costly fingerprinting, and it is an easy step to regard the work as comparable. But closer inspection of the past work reveals that the groups were all working on different problems.

- Manber (1994) used fingerprints to find similar files on a filesystem. Datasets used were compilations of online documentation such as README files. Documents were distinguished as being "similar" if the proportion of identical fingerprints between the documents exceeded a threshold, for example 50%. Manber reports the number of clusters of "similar" documents found by his technique, but does not report on any investigation of the nature of the similarities found by the system.
- Brin et al. (1995) investigated fingerprinting in the context of copyright protection in digital libraries. The dataset used for experimentation was a small collection of academic articles. These articles were manually grouped into "related" documents and the scores between these were compared to the scores between unrelated documents. The conclusion was that there was a large difference between the scores.
- Heintze (1996) investigated the characteristics of different fingerprinting schemes. The dataset was a small collection of technical reports. The experiments compare various fingerprint selection schemes with full fingerprinting, in which every fingerprint is stored. The findings are that sensitivity of the algorithm is not heavily affected by increasing the selectivity of fingerprint selection.
- Broder (1997) used fingerprinting to find documents that are "roughly the same", based on *resemblance* and *containment*, defined by a count of the volume of text two documents share. The motivation is management of web data. The dataset was a large crawl of web documents. Results focused on the runtime of the algorithm, with a brief mention of the number of identical and "similar" documents found.
- Chowdhury et al. (2002) identify documents that are identical after removal of common terms. The motivation is improving search-engine performance. The datasets used are a set of web documents from the Excite@Home crawl thought to have duplicates within the collection, a subset of the TREC LATimes collection with known duplicates seeded into the collection, TREC disks 4 and 5, and WT2G. Synthetic "duplicates" were created by permuting existing documents. Success was measured by the proportion of known duplicates discovered by various methods.
- Fetterly et al. (2003) used a variant of fingerprinting known as super-shingling to analyze large web collections for "near-duplicates" with a 90% likelihood of two fingerprints matching between documents that are 95% similar. Similarity is defined by whether the fingerprints match. The motivation is improved crawling. The results were that they found large numbers of clusters of documents that shared fingerprints.
- Our work (Bernstein and Zobel, 2004) concerned detection of co-derivative documents, that is, documents that were derived from each other or from some other document. We used a test collection composed of documentation from distributions of RedHat Linux, and verified the detected duplicates for a sample of query documents. Measures were analogous to recall and precision. Experimental findings were that our technique was reasonably accurate at finding co-derived documents.

There are good reasons to want to identify duplicates. They may represent redundant information; intuitively, there seems no reason to store the same information multiple

times, and it is rarely helpful to have multiple copies of a document in an answer list. Elimination of duplicates may have benefits for efficiency at search time. In a web collection, the presence of duplicates can indicate a crawler failure. Knowledge of duplication can be used for version management or file system management, and can plausibly be used to help identify where an item of information originated (Metzler et al., 2005). And copies of information may be illegitimate.

However, in much of the prior work in the area, the different kinds of duplication, and the different ways in which knowledge of duplication might be used, were jumbled together. There was no consideration of whether the information about duplicates could be used to solve a practical problem and, fundamentally, in none of these papers was there a qualitative definition of what a duplicate was. Without such a definition, it is not clear how the performance of these systems might be measured, or how we could evaluate whether they were doing useful work. Over the next few sections we explore the difficulties of measurement in the context of research, then return to the question of duplicate detection.

## 3 Research and Measurement

Successful research leads to change in the practice or beliefs of others. We persuade people to use a new algorithm, or show that an existing belief is wrong, or show how new results might be achieved, or demonstrate that a particular approach is effective in practice. That is, research is valuable if the results have impact and predictive power. Research is typically pursued for subjective or context-dependent reasons – for example, we find the topic interesting or look into it because we have funding for investigation of a certain problem.

However, research outcomes are expected to be objective, that is, free from the biases and opinions of the researcher doing the work. If a hypothesis is objectively shown to be false, then it is false, no matter how widely it is believed or how true it had seemed to be; and, if there is solid evidence to support a hypothesis, then probably it should be believed, even if it seems to contradict intuition. That is, we say the hypothesis is *confirmed*, meaning that the strength of belief in the hypothesis is increased.

For research to be robust and to have high impact, three key elements must be present. First, the hypothesis being investigated must be interesting – that is, if it is confirmed, then it will alter the practice and research of others. Second, there must be a convincing way of *measuring* the outcomes of the research investigation. Third, according to this measure the hypothesis should be confirmed. In this paper, we call the thing being measured a *system* and the measure a *yardstick*. Examples of systems include a search engine, a sorting algorithm, and a web crawler; these are bodies of code that have identifiable inputs and are expected to produce output meeting certain criteria. Examples of yardsticks include computation time on some task, number of relevant documents retrieved, and time for a human to complete a task using a system.

Without measurement, there are no research outcomes. Nothing is learnt until a measurement is taken. The onus is on the researcher to use solid evidence to persuade a skeptical reader that the results are sound; how convincing the results are will partly depend on how they are measured. “A major difference between a ‘well-developed’ sci-

ence such as physics and some of the less 'well-developed' sciences such as psychology or sociology is the degree to which things are measured" (Roberts, 1979, page 1).

How a system is measured is a choice made by the researcher. It is a subjective choice, dictated by the expected task for which the system will be used or the expected context of the system. For example, will the system be run on a supercomputer or a palmtop? Will the user be a child or an expert? Will the data to be searched be web pages or textbooks? There is no authority that determines what the yardstick for any system should be. For measurement of a research outcome such as an interface, this observation is obvious; what may be less obvious is that the observation also applies to algorithmic research.

Consider empirical measurement of the efficiency of some algorithm whose properties are well understood, such as a method for sorting integers. The efficiency of an algorithm is an absolutely fundamental computer science question, but there are many different ways to measure it. We have to choose test data and specify its properties. We then have to make assumptions about the environment, such as the volume of data in relation to cache and memory and the relative costs of disk, network, processor, and memory type. There is no absolute reference that determines what is a reasonable "typical" amount of buffer memory for a disk-based algorithm should be, or whether an algorithm that uses two megabytes of memory to access a gigabyte of disk is in any meaningful way superior to one that is faster but uses three megabytes of memory.

Complexity, or asymptotic cost, is widely accepted as a measurement of algorithmic behaviour. Complexity can provide a clear reason to choose one algorithm over another, but it has significant limitations as a yardstick. To begin with, "theoretical results cannot tell the full story about real-world algorithmic performance" (Johnson, 2002). For example, the notional cost of search of a B-tree of $n$ items is $O(\log n)$, but in practice the cost is dominated by the effort of retrieval of a single leaf node from disk. A particular concern from the perspective of measurement is that complexity analysis is based on subjective decisions, because it relies on assumptions about machine behaviour and data. Worst cases may be absurd in practice; there may be assumptions such as that all memory accesses are of equal cost; and average cases are often based on simplistic models of data distributions. Such issues arise in even elementary algorithms. In an in-memory chained hash table, for example, implemented on a 2005 desktop computer, increasing the number of slots decreases the per-slot load factor – but can increase the per-key access time for practical data volumes (Askitis and Zobel, 2005).

While a complexity analysis can provide insight into behaviour, such as in comparison of radixsort to primitive methods such as bubblesort, it does not follow that such analysis is always sufficient. First, "only experiments test theories" (Tichy, 1998). Second, analysis is based on assumptions as subjective as those of an experiment; it may provide no useful estimate of cost in practice; and it is not the answer to the problem of the subjectivity of measurement.

Philosophical issues such as paradoxes of measurement are not merely academic concerns, but are significant practical problems in design of research projects. We need to find a basis for justification of our claims about research outcomes, to guide our work and to yield results that are supported by plausible, robust evidence.

## 4 Choosing a Yardstick

Identification of what to measure is a key step in development of an idea into a concrete research project. In applied science, the ultimate aim is to demonstrate that a proposal has *utility*. The two key questions are, thus, what aspect of utility to measure and how to measure it. We propose that principles for choice of a process of measurement – that is, choice of yardstick – be based on the concept of a *warrant*. Booth et al. (1995) define a warrant as an assumption that allows a particular kind of evidence to be used to support a particular class of hypothesis. An example from Booth et al. is:

**Hypothesis.** It rained last night.
**Evidence.** The streets are wet this morning.

This argument may seem self-supporting and self-evident. However, the argument relies on an implied warrant: that the most likely cause of wet streets is rain. Without the warrant, there is nothing to link the evidence to the hypothesis. Crucially, there is nothing within either the hypothesis or the evidence that is able to justify the choice of warrant; the warrant is an assertion that is external to the system under examination.

The fact that the warrants under which an experiment are conducted are axiomatic can lead to a kind of scientific pessimism, in which results have no authority because they are built on arbitrary foundations. With no criteria for choosing amongst warrants, we are in the position of the philosopher who concludes that all truths are equally likely, and thus that nothing can be learnt. However, clearly this is unhelpful: some warrants do have more merit than others. The issue then becomes identification of the properties a good set of warrants should have.

The answer to the question "what should we measure?" we refer to as the qualitative warrant, and the answer to the question "how should we measure it?" we refer to as the quantitative warrant, that is, the yardstick. These assertions are what links the measurement to the goal of demonstrating utility. We propose a set (not necessarily exhaustive) of four properties that are satisfied by a good qualitative warrant, and of three properties that are satisfied by a good yardstick:

- **Applicability.** A qualitative warrant should reflect the task or problem the system is designed to address. For example, it would (usually) be uninteresting to measure a user interface based on the number of system calls required to render it.
- **Power.** The power of a qualitative warrant is the degree to which it makes a meaningful assertion about utility. Intuitively, a qualitative warrant is not powerful if its negation results in a new warrant that seems equally reasonable. For example, the warrant "a system is useful if it discards documents that are of uncertain relevance" is not powerful, because its negation, "a system is useful if it retains documents that are of uncertain relevance", also seems reasonable. In contrast, the warrant "an algorithm is useful if it can sort integers faster than any known algorithm" is powerful because its negation, "an algorithm is useful if it cannot sort integers faster than other algorithms", is absurd.
- **Specificity.** Evaluation of a system cannot be meaningful if we are not specific about what we are trying to measure. An example is a warrant such as "a system is useful if it allows users quick access to commonly-used functions". While at first

glance this may seem reasonable, the question of which functions are commonly used is likely to depend on the task and the kind of user.
– **Richness.** The utility of many systems depends on more than just one dimension of performance. For example, we would like an information retrieval system to be both fast and effective. The speed of an IR system can be the basis of a qualitative warrant that is both applicable and poweful; however, it misses a key aspect of IR system performance. Hence, we say that the warrant lacks richness.

The quantitative warrant is effectively dictated by the choice of yardstick used to measure the system. A good yardstick should have the following properties:

– **Independence.** A yardstick needs to be independent of the solution; it should not rely in a circular way on the system being measured, but should instead be defined in terms of some properties that would still be meaningful even if the system did not exist. If we claim that a method is useful because it finds grue documents, and that documents are grue if they are found by the method, then the "grueness" yardstick is meaningless. Ethical issues are also relevant; a researcher should not, for example, choose a yardstick solely on the basis that it favours a particular system.
– **Fidelity.** Because the yardstick is used to quantify the utility of the system under investigation, there needs to be fidelity, that is, a strong correspondence between the outcome as determined by the yardstick and the utility criterion it is attempting to quantify. Many yardsticks must reduce a complex process to a simple quantifiable model, that is, "most representations in a scientific context result in some reduction of the original structure" (Suppes et al., 1994). Success by a yardstick lacking fidelity will not be meaningful outside the context of the research.
– **Repeatability.** We expect research results to be predictive, and in particular that repeating an experiment will lead to the same outcomes. The outcomes may vary in detail (consider a user experiment, or variations in performance due to machines and implementation) but the broad picture should be the same. Thus the yardstick should measure the system, not other factors that are external to the work.

Using these criteria, it can be argued that some qualitative warrants are indeed superior to others, and that, given a particular qualitative warrant, some yardsticks are superior to others. Note that measures often conflict, and that this is to be expected – consider yardsticks such as speed versus space, or speed versus complexity of implementation, or speed in practice versus expected asymptotic cost. We should not expect yardsticks to be consistent, and indeed this is why choice of yardstick can be far from straightforward.

For algorithmic work, we may choose a qualitative warrant such as "an algorithm is useful if it is computationally efficient". This satisfies the criteria: it is applicable, powerful, reasonably specific, and rich. Given this warrant, we can consider the yardstick "reduced elapsed computation time". It is independent (we don't even need to know what the algorithm is), repeatable, and in general is a faithful measure of utility as defined by the qualitative warrant. The yardstick "reduced instruction count" is independent and repeatable, but in some cases lacks fidelity: for many algorithms, other costs, such as memory or disk accesses, are much more significant. The yardstick "makes use of a wider range of instructions" is independent and repeatable, but entirely lacks fidelity: measures by this yardstick will bear little correspondence to utility as defined by our qualitative warrant.

Some potential criteria that could be used to justify a yardstick are fallacies or irrelevancies that do not stand scrutiny. For example, the fact that a property is easy to measure does not make the measure a good choice. A yardstick that has been used for another task may well be applicable, but the fact that it has been used for another task carries little weight by itself; the rationale that led to it being used for that task may be relevant, however. Even the fact that a yardstick has previously been used for the same task may carry little weight – we need to be persuaded that the yardstick was well chosen in the first place.

An underlying issue is that typically yardsticks are abstractions of semantic properties that are inherently not available by symbolic reasoning. When a survey is used to measure human behaviour, for example, a complex range of real-world properties is reduced to numerical scores. Confusion over whether processes are "semantic" is a failing of a range of research activities. Symbolic reasoning processes cannot be semantic; only abstract representations of real-world properties – not the properties themselves, in which the meaning resides – are available to computers.

Note too that, as computer scientists, we do not write code merely to produce software, but to create a system that can be measured, and that can be shown to possess a level of utility according to some criterion. If the principal concern is efficiency, then the code needs to be written with great care, in an appropriate language; if the concern is whether the task is feasible, a rapid prototype may be a better choice; if only one component of a system is to be measured, the others may not need to be implemented at all. Choice of a yardstick determines which aspects of the system are of interest and thus need to be implemented.

## 5 Measurement in Information Retrieval

In algorithmic research, the qualitative warrants are fairly straightforward, typically concerning concrete properties such as speed, throughput, and correctness. Such warrants can be justified – although usually the justification is implicit – by reference to goals such as reducing costs. Yardsticks for such criteria are usually straightforward, as the qualitative warrants are inherently quantifiable properties.

In IR, the qualitative warrant concerns the quality of the user search experience, often in terms of the cost to the user of resolving an information need. Yardsticks are typically based on the abstractions *precision* and *recall*. The qualitative warrant satisfies the criteria of applicability, power, and richness. Furthermore, the IR yardsticks typically demonstrate independence and repeatability.

However, the qualitative warrant is not sufficiently specific. It is difficult to model user behaviour when it has not been specified what sort of user is being modelled, and what sort of task they are supposed to be performing. For example, a casual web searcher does not search in the same way as a legal researcher hoping to find relevant precedents for an important case. Even if the qualitative warrant were made more specific, for example by restricting the domain to ad-hoc web search, the fidelity of many of the current yardsticks can be brought into question. Search is a complex cognitive process, and many factors influence the degree of satisfaction a user has with their search experience; many of these factors are simplified or ignored in order to yield a yard-

stick that can be tractably evaluated. It is not necessarily the case that the user will be most satisfied with a search that simply presents them with the greatest concentration of relevant documents.

To the credit of the IR research community, measurement of effectiveness has been the subject of ongoing debate; in some other research areas, the issue of measurement is never considered. In particular, user studies have found some degree of correlation between these measures and the ease with which users can complete an IR task (Allan et al., 2005), thus demonstrating that – despite the concerns raised above – the yardsticks have at least limited fidelity and research outcomes are not entirely irrelevant.

Yardsticks drive the direction of research; for example, the aim of a great deal of IR research is to improve recall and precision. To the extent that a yardstick represents community agreement on what outcome is desirable, letting research follow a yardstick is not necessarily a bad thing. However, if the divergence between yardsticks and the fundamental aim of the research is too great, then research can be driven in a direction that is not sensible. We need to be confident that our yardsticks are meaningful in the world external to the research.

## 6 Measurement of Duplicate Discovery Methods

In some of the work on duplicate discovery discussed earlier, the qualitative warrant is defined as (we paraphrase) "system $A$ is useful if it is able to efficiently identify duplicates or near-duplicates". However, anything that is found by the algorithms is deemed to be a duplicate. Such a yardstick clearly fails the criteria set out earlier. It is not independent, powerful, or rich. It provides no guidance for future work, or any information as to whether the methods are valuable in practice.

The question of whether these methods are successful depends on the definition of "duplicate". When the same page is crawled twice, identical but for a date, there are still contexts in which the two versions are not duplicates – sometimes, for example, the dates over which a document existed are of interest. Indeed, almost any aspect of a document is a reasonable target of a user's interest. It is arguable whether two documents are duplicates if the name of the author has changed, or if the URL is different. Are a pair "the same" if one byte is changed? Two bytes? That is, there is no one obvious criterion for determining duplication. Again, we argue that the warrant is not specific enough. A pair of documents that are duplicates in the context of, say, topic search may not be duplicates in the context of, say, establishing which version is most up-to-date.

As in IR, there need to be clear criteria against which the assessment is made, in which the concept of task or utility is implicitly or explicitly present. For duplication, one context in which task can be defined is that of search. Consider some of the ways in which a document might address an information need:

- As a source of new information.
- As confirmation of existing knowledge.
- As a means of identifying the author, date, or context.
- As a demonstration of whether the information is from a reputable provider.

That is, a pair of documents can only be judged as duplicates in the context of the use that is being made of them.

To establish whether our SPEX method for duplicate discovery was effective, we explored several search-oriented varieties of duplication, using human experiments to calibrate SPEX scores against human judgements (Bernstein and Zobel, 2005).

The first kind of duplication was *retrieval equivalence*: a pair of documents is retrieval equivalent if they appear identical to a typical search engine. This definition can be validated mechanically, by parsing the documents according to typical search engine rules to yield a canonical form. A pair is deemed to be retrieval equivalent if their canonical forms are bytewise identical. However, even retrieval equivalent documents may not be duplicates in the context of some search tasks. Two mirrors may hold identical documents, but we may trust one mirror more than another; removal of either document from an index would be a mistake. Knowledge of duplication can affect how such answers are presented, but does not mean that they can be eliminated.

The second kind of duplication we considered was *content equivalence*. In an initial experiment, we identified document pairs where SPEX had returned a high match score, and asked test subjects to assess the pairs against criteria such as "effectively duplicated". However, our subjects differed widely in their interpretation of this criterion. For some, a minor element such as date was held to indicate a substantive difference; for others it was irrelevant. We therefore refined these criteria, to statements such as "differences between the documents are trivial and do not differentiate them with respect to any reasonable query" and "with respect to any query for which both documents may be returned by a plausible search, the documents are equivalent; any query for which the documents are not equivalent would only return one or the other". We called this new criterion *conditional equivalence*.

We could define our warrants for this task as follows:

**Qualitative warrant.** The SPEX system is useful if it accurately identifies pairs of documents that can be considered to be duplicates in a web search context.

**Quantitative warrant.** The extent to which pairs of documents identified by a system are judged by a human to be duplicates in a web search context is a good estimator of whether the system accurately identifies duplicates.

Superficially, retrieval and content equivalence, and the sub-classes of content equivalence, may seem similar to each other, but in a good fraction of cases documents that were duplicates under one criterion were not duplicates under another. An immediate lesson is that investigation of duplicate discovery that is not based on a clear definition of task is meaningless. A more positive lesson is that these definitions provide a good yardstick; they meet all of the criteria listed earlier.

Using these yardsticks, we observed that there was a clear correlation between SPEX scores and whether a user would judge the documents to be duplicated. This meant that we could use SPEX to measure the level of duplication – from the perspective of search! – in test collections. Our experiments used the GOV1 and GOV2 collections, two crawls of the `.gov` domain created for TREC. GOV1 is a partial crawl of `.gov` from 2002, with 1,247,753 documents occupying 18 gigabytes. GOV2 is a much more complete crawl of `.gov` from 2004, with 25,205,179 documents occupying 426 gigabytes.

On the GOV1 collection, we found that 99,227 documents were in 22,870 retrieval-equivalent clusters. We found a further 116,087 documents that participated in content-equivalence relationships, and that the change in definition from content-equivalence

to conditional equivalence led to large variations in the numbers of detected duplicates. On the GOV2 collection, we found a total of 6,943,000 documents in 865,362 retrieval-equivalent clusters – more than 25% of the entire collection. (Note that, prior to distribution of the data, 2,950,950 documents were removed after being identified as duplicates by MD5.) Though we were unable to scan the entire GOV2 collection for content-equivalence, we believe that a similar proportion again is content-equivalent, as was the case for the GOV1 collection.

These results indicate that there are many pairs of documents within these collections that are mutually redundant from a user perspective: if a user were to see one document in relation to a particular query, there may be several other documents that would no longer be of interest to them. This observation provides empirical support to the questioning of the notion of independent relevance. The results suggest that the volume of retrieval- and content-equivalent documents in the collection may be so great that the assumption of independent relevance is significantly affecting the fidelity of the IR yardsticks.

To investigate this further, we experimented with the runs submitted for the TREC 2004 terabyte track, consisting of result sets for 50 queries on the GOV2 collection. In our first experiment, we modified the query relevance assessments so that a document returned for a particular query on a particular system would be marked as not relevant if a document with which it was content-equivalent appeared earlier in the result list. This partially models the notion of relevance as dependent on previously seen documents. The result was significant: under this assumption, the MAP of the runs in the top three quartiles of submission dropped by a relative 20.2% from 0.201 to 0.161. Interestingly, the drop in MAP was greater for the more successful runs than for the less successful runs. While ordering between runs was generally preserved, it seems that the highest-scoring runs were magnifying their success by retrieving multiple copies of the same relevant document, an operation that we argue does nothing to improve the user search experience in most cases.

These experiments allowed us to observe the power that measurement and yardsticks have in influencing the direction of research. Consider two examples.

The first example is that, in defining an appropriate measure of the success of our system, we were forced to re-evaluate and ultimately redefine our task. We had originally intended to simply measure the occurrence in collections of documents that were content-equivalent with a view to removing them from the collection. Our user experiments quickly showed us that this approach was unrealistic: even minor differences between documents had the potential to be significant in certain circumstances. The concept of conditional equivalence, in which documents were equivalent with respect to a query, proved to be far more successful. This meant that it was unsuitable to simply remove documents from the collection; rather, duplicate removal was much better performed as a postprocessing step on result lists. This lesson, learnt in the process of defining a yardstick, has practical effects on the way in which duplication should be managed in search engines.

The second example concerns the fidelity of measures based on the assumption of independence of relevance. We have shown that, based on user experiments, our software can reliably identify pairs of documents that are conditionally equivalent, and that

lifting the general assumption of independent relevance can have a significant impact on the reported effectiveness of real search systems. Furthermore, postprocessing result lists in order to remove such equivalent documents, while significantly increasing MAP from the lower figure, failed to restore the MAP of runs to its original level. The consequence of this is that the current TREC assessment regime discourages the removal of duplicate documents from result lists. This demonstrates the power of yardsticks, and the dangers if they are poorly chosen. Because yardsticks are the measured outcomes of research, it is natural for research communities to have as their goal improvement in performance according to commonly accepted yardsticks. Given an insufficiently faithful yardstick it is likely, or perhaps inevitable, that the research activity may diverge from the practical goals that the research community had originally intended to service.

## 7 Conclusions

Careful consideration of how outcomes are to be measured is a critical component of high-quality research. No researcher, one presumes, would pursue a project with the expectation that it will have little impact, yet much research is unpersuasive and for that reason is likely to be ignored. Each paper needs a robust argument to demonstrate that the claims are confirmed. Such argument rests on evidence, and, in the case of experimental research, the evidence depends on a system of measurement.

We have proposed seven criteria that should be considered when deciding how research outcomes should be measured. These criteria – applicability, power, specificity, richness, independence, fidelity, and repeatability – can be used to examine yardsticks used for measurement. As we have argued in the case of IR research, even widely accepted yardsticks can be unsatisfactory. In the case of the duplicate documents, our examination of the problems of measurement reveals one plausible reason why some prior work has had little impact: the yardsticks are poor or absent, and consequently the work is not well founded.

We applied the criteria to evaluation of our new yardsticks for duplicate detection, and found that the concept of "duplicate" is surprisingly hard to define, and in the absence of a task is not meaningful. Almost every paper on duplication concerns a different variant and our user studies show that slightly different definitions of "duplicate" lead to very different results. Duplicates can be found, but there is no obvious way to find specific kinds of duplicates – previous work was typically motivated by one kind of duplication but measured on all kinds of duplication. Our examination of yardsticks not only suggests future directions for research on duplicate detection, but more broadly suggests processes that researchers should follow in design of research projects.

*Acknowledgements.* This work was supported by the Australian Research Council.

## References

Allan, J., Carterette, B. and Lewis, J. (2005), When will information retrieval be "good enough"?, *in* "Proc. ACM-SIGIR Ann. Int. Conf. on Research and Development in Information Retrieval", ACM Press, New York, NY, USA, pp. 433–440.

Askitis, N. and Zobel, J. (2005), Cache-conscious collision resolution in string hash tables, *in* "Proc. String Processing and Information Retrieval Symposium (SPIRE)". To appear.

Bernstein, Y. and Zobel, J. (2004), A scalable system for identifying co-derivative documents, *in* A. Apostolico and M. Melucci, eds, "Proc. String Processing and Information Retrieval Symposium (SPIRE)", Springer, Padova, Italy, pp. 55–67.

Bernstein, Y. and Zobel, J. (2005), Redundant documents and search effectiveness, *in* "Proc. ACM Ann. Int. Conf. on Information and Knowledge Management (CIKM)". To appear.

Booth, W. C., Colomb, G. G. and Williams, J. M. (1995), *The Craft of Research*, U. Chicago Press.

Brin, S., Davis, J. and García-Molina, H. (1995), Copy detection mechanisms for digital documents, *in* M. Carey and D. Schneider, eds, "Proc. ACM-SIGMOD Ann. Int. Conf. on Management of Data", ACM Press, San Jose, California, United States, pp. 398–409.

Broder, A. Z. (1997), On the resemblance and containment of documents, *in* "Compression and Complexity of Sequences (SEQUENCES'97)", IEEE Computer Society Press, Positano, Italy, pp. 21–29.

Chowdhury, A., Frieder, O., Grossman, D. and McCabe, M. C. (2002), "Collection statistics for fast duplicate document detection", *ACM Transactions on Information Systems (TOIS)* **20**(2), 171–191.

Fetterly, D., Manasse, M. and Najork, M. (2003), On the evolution of clusters of near-duplicate web pages, *in* R. Baeza-Yates, ed., "Proc. 1st Latin American Web Congress", IEEE, Santiago, Chile, pp. 37–45.

Heintze, N. (1996), Scalable document fingerprinting, *in* "1996 USENIX Workshop on Electronic Commerce", Oakland, California, USA, pp. 191–200.

Johnson, D. S. (2002), A theoretician's guide to the experimental analysis of algorithms, *in* M. Goldwasser, D. S. Johnson and C. C. McGeoch, eds, "Proceedings of the 5th and 6th DIMACS Implementation Challenges", American Mathematical Society, Providence.

Manber, U. (1994), Finding similar files in a large file system, *in* "Proc. USENIX Winter 1994 Technical Conference", San Fransisco, CA, USA, pp. 1–10.

Metzler, D., Bernstein, Y., Croft, W. B., Moffat, A. and Zobel, J. (2005), Similarity measures for tracking information flow, *in* "Proc. ACM Ann. Int. Conf. on Information and Knowledge Management (CIKM)". To appear.

Moffat, A. and Zobel, J. (2004), What does it mean to 'measure performance'?, *in* X. Zhou, S. Su, M. P. Papazoglou, M. E. Owlowska and K. Jeffrey, eds, "Proc. International Conference on Web Informations Systems", Springer, Brisbane, Australia, pp. 1–12. Published as LNCS 3306.

Roberts, F. S. (1979), *Measurement Theory*, Addison-Wesley.

Suppes, P., Pavel, M. and Falmagne, J.-C. (1994), "Representations and models in psychology", *Annual Review of Psychology* **45**, 517–544.

Tichy, W. F. (1998), "Should computer scientists experiment more?", *IEEE Computer* **31**(5), 32–40.