On Measurement

Justin Zobel

University of Melbourne, Australia

Semester 2, 2017

Measurement

What is to be measured? What measures are to be used?

What should be used as a baseline? If the measurements improve on the baseline, what does that show?

How do the measured properties relate to the aims of the research?

Measures can be categorical (yes/no, was/was not); by rank; by value.

Measurement is used to develop understanding of a system or to test some predictions, that is, test whether some predicted behaviour does in fact occur. Distinguish these:

- Exploration: collect data, analyse it, search for pattern.
- Evaluation: measure the effect of changes on system behaviour, compare to predictions.

Example: Document collections and duplication

There are good reasons to want to identify duplicates:

- Duplicates may represent redundant information. Should the same information be stored multiple times?
- It is rarely helpful to have multiple copies of a document in an answer list.
- 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.
- Knowledge of duplication can plausibly be used to help identify where an item of information originated.
- Copies of information may be illegitimate.

What is a duplicate?

In some of the research on duplicate detection, the task is defined as (paraphrased) ‘identification of duplicates or near-duplicates’. Anything that is found by the algorithms is deemed to be a duplicate.

But what sort of duplicates are found?

Are we being misled by the word ‘duplicate’?

- It *seems* to have a simple natural interpretation.
- But is there in fact any agreement of what a duplicate is, or what to do with one when it is found?
- How do we tell whether two documents are duplicates?

What is being measured?

What is a duplicate? . . .

There are many sources of duplicates in typical collections.

- Mirrors.
- Crawl artifacts, such as the same text with a different date or a different advertisement, available at multiple URLs.
- Versions created for different delivery mechanisms.
- Annotated and unannotated copies of documents.
- Jurisdictional variations, such as 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 of text (legitimate and otherwise).

What is a duplicate? …

After how much change is a version no longer a duplicate?

- One bit? One byte? One word?
- When the appearance has changed? (Consider a document rendered in different browsers.)
- When the contents have a semantically significant difference?

Does markup matter?

Does the URL matter?

Are revision numbers or dates important?

It’s science, right?

If this is experimental research that is expected to form a scientific contribution and have impact, we would expect to find the following elements:

- An interesting hypothesis concerning some testable ideas.
- Experiments on implementations of these ideas.
- Objective measurement of the outcomes of the experiments.

That is, there should be a hypothesis and some persuasive evidence that confirms or disproves it.

Objectivity is crucial. We are unlikely to be persuaded by evidence based on the researcher’s personal preference.

Grueishness

In a hypothetical research project, a young scientist, Sherman, is interested in finding documents that are *grue*. This is a new property that Sherman has chosen to investigate.

It is interesting, he says, because grue documents tend to be bleen.

Grueishness

In a hypothetical research project, a young scientist, Sherman, is interested in finding documents that are *grue*. This is a new property that Sherman has chosen to investigate.

It is interesting, he says, because grue documents tend to be bleen.

Sherman devises a heuristic algorithm for locating grue documents.

Grueishness

In a hypothetical research project, a young scientist, Sherman, is interested in finding documents that are *grue*. This is a new property that Sherman has chosen to investigate.

It is interesting, he says, because grue documents tend to be bleen.

Sherman devises a heuristic algorithm for locating grue documents.

A document is defined as grue if it is found by the algorithm.

Grueishness

In a hypothetical research project, a young scientist, Sherman, is interested in finding documents that are *grue*. This is a new property that Sherman has chosen to investigate.

It is interesting, he says, because grue documents tend to be bleen.

Sherman devises a heuristic algorithm for locating grue documents.

A document is defined as grue if it is found by the algorithm.

- Which does not seem like a very interesting research outcome. The method ‘works’ by definition.

Grueishness

In a hypothetical research project, a young scientist, Sherman, is interested in finding documents that are *grue*. This is a new property that Sherman has chosen to investigate.

It is interesting, he says, because grue documents tend to be bleen.

Sherman devises a heuristic algorithm for locating grue documents.

A document is defined as grue if it is found by the algorithm.

- Which does not seem like a very interesting research outcome. The method ‘works’ by definition.

Surely real science by experienced, competent researchers isn’t like this . . . is it?

Duplicates?

Meaningful research or meaningless grue?

- Broder (1997) reported the number of high-scoring document pairs found by his algorithm.
- Chowdhury, Frieder, Grossman, and McCabe (2002) found that different heuristics identified different numbers of duplicates.
- Fetterly, Manasse, and Najork (2003) found large numbers of documents that scored highly according to their metric. They found a 90% likelihood of two fingerprints matching between documents that are 95% similar (identical documents with 5% random corruption).
- Bernstein and Zobel (2004) found that their metrics discovered most of the documents that shared a significant amount of text.

This is science

Successful research leads to change in the practice and beliefs of others.

In principle, good science uses objective evidence to achieve aims such as:

- Persuade us to adopt a new algorithm.
- Show us that an existing belief is wrong.
- Help us to choose between methods that we might apply in practice.

Presumably, all scientists want their work to be persuasive!

Is this science?

In the area of duplicate detection, something is not right.

- There is no agreement – or even discussion! – on what to measure.

That is, there is no *qualitative*, in-principle aim: What are the properties that the algorithm is intended to discover?

- There is no proposal for a method of measurement.

That is, there is no *yardstick*, other than the algorithm itself.

If the yardstick is to be objective, we need an objective way of choosing it.

An easy case?

Sherman is working on a search algorithm. He believes he has found a way to improve the work of his colleague, Mr. Peabody.

> SHERMAN: *My algorithm is faster than yours.*

An easy case?

Sherman is working on a search algorithm. He believes he has found a way to improve the work of his colleague, Mr. Peabody.

> SHERMAN: *My algorithm is faster than yours.*
>
> MR. PEABODY: *But it uses more memory.*

An easy case?

Sherman is working on a search algorithm. He believes he has found a way to improve the work of his colleague, Mr. Peabody.

> Sherman: *My algorithm is faster than yours.*
>
> Mr. Peabody: *But it uses more memory.*
>
> Sherman: *But the extra memory is a constant overhead, while the algorithm is five times faster.*

An easy case?

Sherman is working on a search algorithm. He believes he has found a way to improve the work of his colleague, Mr. Peabody.

> Sherman: *My algorithm is faster than yours.*
>
> Mr. Peabody: *But it uses more memory.*
>
> Sherman: *But the extra memory is a constant overhead, while the algorithm is five times faster.*
>
> Mr. Peabody: *You tested your method on an ARM processor. Your algorithm would not work well on a Pentium. And it would still use more memory.*

An easy case?

Sherman is working on a search algorithm. He believes he has found a way to improve the work of his colleague, Mr. Peabody.

> SHERMAN: *My algorithm is faster than yours.*
>
> MR. PEABODY: *But it uses more memory.*
>
> SHERMAN: *But the extra memory is a constant overhead, while the algorithm is five times faster.*
>
> MR. PEABODY: *You tested your method on an ARM processor. Your algorithm would not work well on a Pentium. And it would still use more memory.*

Who is right – Mr. Peabody or Sherman?

It seems that subjective choice (in this case, of computing platform) needs to be part of the evidence.

Utility and subjectivity

The ultimate aim of applied science is to demonstrate that a proposal has practical *utility*.

This is obviously true of disciplines such as medicine and architecture, and, within computer science, areas such as HCI.

Less obviously, it is true across much of computer science.

Utility has parallels with the information retrieval concept of relevance: both require a human assessment.

An immediate consequence of reliance on utility as a benchmark of scientific outcomes is a paradox of measurement.

- Evidence is supposed to be *objectively* measured.
- How we measure evidence is a *subjective* choice, based entirely on human assessment of utility.

Weights and measures

Ken Adler, *The Measure of All Things*:

- ‘Under the cover of some eight hundred names, *ancien régime* France contained a staggering 250,000 different units of weights and measures.’
- ‘Many *ancien régime* measures . . . derived from human needs and human interests [and] reflected the quantity of labour a person could do in a given period of time . . . surface area [measurements] would actually vary depending on the type of field and the quality of its soil.’
- ‘The technique of measurement depended on local custom. One district measured grain heaped high in its bushel; another measured grain after it had been levelled off; still another, after the bushel had been struck to settle its contents.’

Back in the lab

Sherman is still arguing for his search algorithm.

Back in the lab

Sherman is still arguing for his search algorithm.

> SHERMAN: *My search algorithm has excellent recall. For this query with 7 relevant documents, they all turn up in the top 12. For this query with 80 relevant documents, they all turn up in the top 400.*

Back in the lab

Sherman is still arguing for his search algorithm.

> Sherman: *My search algorithm has excellent recall. For this query with 7 relevant documents, they all turn up in the top 12. For this query with 80 relevant documents, they all turn up in the top 400.*
>
> Mr. Peabody: *But it doesn't put the best answers first. Won't the user give up quickly?*

Back in the lab

Sherman is still arguing for his search algorithm.

> Sherman: *My search algorithm has excellent recall. For this query with 7 relevant documents, they all turn up in the top 12. For this query with 80 relevant documents, they all turn up in the top 400.*
>
> Mr. Peabody: *But it doesn't put the best answers first. Won't the user give up quickly?*
>
> Sherman: *An experienced user will keep looking until they have seen all the answers.*

Back in the lab

Sherman is still arguing for his search algorithm.

> Sherman: *My search algorithm has excellent recall. For this query with 7 relevant documents, they all turn up in the top 12. For this query with 80 relevant documents, they all turn up in the top 400.*
>
> Mr. Peabody: *But it doesn't put the best answers first. Won't the user give up quickly?*
>
> Sherman: *An experienced user will keep looking until they have seen all the answers.*
>
> Mr. Peabody: *But how does the user know when to stop looking?*

Back in the lab

Sherman is still arguing for his search algorithm.

> Sherman: *My search algorithm has excellent recall. For this query with 7 relevant documents, they all turn up in the top 12. For this query with 80 relevant documents, they all turn up in the top 400.*
>
> Mr. Peabody: *But it doesn't put the best answers first. Won't the user give up quickly?*
>
> Sherman: *An experienced user will keep looking until they have seen all the answers.*
>
> Mr. Peabody: *But how does the user know when to stop looking?*

'There is no absolute sense in which one can say that one particular pair of precision-recall values is better or worse than some other pair.' (van Rijsbergen, 1979)

In the lab again

Sherman now presents what he believes is a decisive point.

In the lab again

Sherman now presents what he believes is a decisive point.

> SHERMAN: *My algorithm has better worst-case asymptotic cost than yours. It is* $O(n \log n)$*, while your algorithm has worst case* $O(n^2)$*.*

In the lab again

Sherman now presents what he believes is a decisive point.

> SHERMAN: *My algorithm has better worst-case asymptotic cost than yours. It is* $O(n \log n)$*, while your algorithm has worst case* $O(n^2)$*.*

> MR. PEABODY: *The worst case is absurdly improbable.*

In the lab again

Sherman now presents what he believes is a decisive point.

> SHERMAN: *My algorithm has better worst-case asymptotic cost than yours. It is* $O(n \log n)$*, while your algorithm has worst case* $O(n^2)$*.*
>
> MR. PEABODY: *The worst case is absurdly improbable.*
>
> SHERMAN: *Anyway, for typical strings, the asymptotic costs of our methods are the same. So my algorithm is at least as good as yours.*

In the lab again

Sherman now presents what he believes is a decisive point.

> SHERMAN: *My algorithm has better worst-case asymptotic cost than yours. It is $O(n \log n)$, while your algorithm has worst case $O(n^2)$.*
>
> MR. PEABODY: *The worst case is absurdly improbable.*
>
> SHERMAN: *Anyway, for typical strings, the asymptotic costs of our methods are the same. So my algorithm is at least as good as yours.*
>
> MR. PEABODY: *Which strings are 'typical'? What about URLs? Lines of code? Names? Words in text? With or without duplicates?*

In the lab again

Sherman now presents what he believes is a decisive point.

> Sherman: *My algorithm has better worst-case asymptotic cost than yours. It is* $O(n \log n)$*, while your algorithm has worst case* $O(n^2)$*.*
>
> Mr. Peabody: *The worst case is absurdly improbable.*
>
> Sherman: *Anyway, for typical strings, the asymptotic costs of our methods are the same. So my algorithm is at least as good as yours.*
>
> Mr. Peabody: *Which strings are 'typical'? What about URLs? Lines of code? Names? Words in text? With or without duplicates?*

Again, an 'objective' research outcome is reduced to a subjective choice.

Complexity as a measure

How useful is complexity as a measure?

- It does not measure cost at all – but only how the cost changes with the data volume.
- Asymptotically dominant costs may be insignificant on any plausible real machine.
- *'Theoretical results cannot tell the full story about real-world algorithmic peformance.'* (Johnson, 2002)
- *'Only experiments test theories.'* (Tichy, 1998)

In any case, complexity analysis is often subjective: it relies on assumptions about machine behaviour and data distribution.

Everything is equal?

If choice of measure is subjective, then where is the science?

- Can a researcher argue that an algorithm is superior because it is slow?

How can we gather any meaningful evidence at all?

We need 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.

Zobel and Bernstein (2006) argue for criteria against which methods of measurement can be assessed.

Assessing qualitative measures

Applicability

A measure should reflect the task the system is designed for.

Consider: Use of instruction counts to measure interfaces?

Power

A measure should be based on meaningful assertions about utility. Intuitively, a measure is not powerful if its negation seems equally plausible.

Consider: Is it better to be slow with low memory usage or fast with high memory usage?

(The concept of power is related to falsifiability.)

Assessing qualitative measures . . .

Specificity

A measure needs to concern the thing being measured.

Consider: Is it reasonable to measure whether a database system (as a whole) is fast? It has many independent components that can contribute to, or confound, the results.

Richness

The utility of many systems depends on more than just one dimension of performance.

Consider: It is enough to assert that a search engine is fast?

Assessing yardsticks

Independence

The yardstick should not be chosen to suit the solution.

Fidelity

Success by the yardstick should correspond to utility.

To take a measurement, it may be necessary to reduce a complex real-world behaviour to a simple quantifiable model.

Consider: How should traffic congestion be measured?

Repeatability

Research results should be predictive of future behaviour.

The yardstick should be such that repeating the experiment leads to the same results (or trends, or relativities).

Defending measures

Consider measurement of an algorithm. A qualitative measure might be:

- ‘An algorithm is useful if it is computationally efficient’.

This is applicable, specific, and rich.

Consider some yardsticks:

- ‘Reduced elapsed computation time’.
  This is independent, faithful, and repeatable.
- ‘Reduced instruction count’.
  Independent and repeatable. Not entirely faithful.
- ‘Makes use of many different instructions’.
  Independent, not particularly repeatable. Not faithful at all.

Relevance revisited

Yardsticks provide direction for research.

In IR, the aim of much work is to improve recall and precision.

- Experiments with relevance, recall, and precision are simple and cheap.
- User experiments are expensive.
- IR outcomes are not always predictive.

However, there has been a great deal of exploration of the relationship between relevance and utility.

Some user experiments suggest that improvements in effectiveness do lead to greater utility. Others do not!

Duplication revisited

For duplication, a qualitative measure might be:

- ‘A duplicate detection system is useful if it is able to identify duplicates or near-duplicates.’

Some duplicates? All duplicates? Is this powerful or rich?

If ‘duplication’ is defined as objects that are found by the system, then the measure lacks independence.

- ‘A duplicate detection system is useful if it is efficient.’

This fails the test of applicability.

Much research is unpersuasive – often because the measures of success are poorly chosen.

Persuasive research stands on careful subjective foundations. Given these foundations, measurement can be objective.