Reverse indexing

David Crystal

The concept of reverse engineering has an interesting theoretical application to indexing, conceiving an index as a means of representing the semantic structure of a book. Indexes only ever achieve partial representation, because of their selectivity, and depend on the notion of relevance. Relevance originates in an indexer's sense of what a book is about. If this sense is lacking, then indexing is problematic. This perspective is applied to the challenge of indexing on the World Wide Web. Maximalist and minimalist approaches are contrasted, and a way of handling the typically multithematic character of web pages is presented.

The index thought experiment

In the beginning was the book, then came the index. But imagine it the other way round. Here's a thought experiment. Imagine a publishing house that holds all its books electronically. A disaster happens, and the indexes are somehow separated from their associated books and all the page numbers wiped out. We are left only with the headwords. Would it be possible to put Humpty back together again? It should be possible. If the indexer has done a thorough job, the index should provide a unique representation of the content. Simply mapping the terms in the index onto the main text of the books should be enough for each index to eventually find its owner.

We can go further. Imagine that the disaster wiped out all the main texts. If the index is strong enough, it ought to be possible to say to a specialist: we don't have the book, but we do have the index. Could you take the index as a brief and write the book that goes with it? It would be time-consuming, and there would be several possible outcomes, but the result would not be bizarre. Indeed, in science such a procedure is well known. It is called 'reverse engineering' – when someone takes a device to pieces to find out how it works, so that another device can be built on similar lines. This would be 'reverse indexing'.

This is, of course, only a thought experiment – a construct of the imagination which helps us investigate the nature of things, like Einstein's elevator or Schrödinger's cat. But thought experiments can have practical outcomes. For what it suggests, in our case, is that the index is a means of generating the semantic content of a book. And to see an index as a representation of the semantic structure of a book is a fruitful notion.

Relevance and partial representation

It can only be a partial representation, of course. The typical alphabetical character of an index obscures many of the semantic relationships it contains. If I am writing a book about kinship relations, I will have aunts under A and uncles under U. We all know that these entities are semantically complementary, but this is hidden by the alphabetical separation, and if we want to make it explicit we have to resort to a convention such as 'see also'. It is a convention that is used sparingly. Every index entry could have a 'see also' reference to some other.

The representation is partial, in real books, also because the selection of entries is governed by such pragmatic principles as usefulness, level of interest, and above all relevance. We are trying to second-guess what readers will want to find. We do not want our entries to be too general or too detailed. And we want readers to feel that our entries are relevant to their concerns.

Relevance is critical. What would happen if we dispensed with it? Let us take the thought experiment a stage further and present the index of a text in which no attention is paid to relevance at all. Here is the opening paragraph from the preface of my most recent book, *By hook or by crook*, which I choose because most of you will not have read it and thus will have no idea what it is about.

> The inspiration for *By hook or by crook* came from reading W. G. Sebald's *The rings of Saturn*, an atmospheric semi-fictional account of a walking tour throughout East Anglia, in which personal reflections, historical allusions and traveller observations randomly combine into a mesmerising novel about change, memory, oblivion and survival. The metaphor of the title – Saturn's rings created from fragments of shattered moons – captures the fragmentary and stream-of-consciousness flow of the narrative.

If we have dispensed with relevance, then we must index everything – for everything is potentially relevant. That would produce a result something like this (I am not concerned with the way these entries are phrased, only with the selection). There are 38 items in the index to this paragraph.

account, of *The rings of Saturn*
allusions, historical
atmosphere, of *The rings of Saturn*
By hook or by crook
capture, of narrative flow
change, nature of
creation, of planetary rings
East Anglia
flow, in narrative

fragmentation, in narrative
fragments, of moons
history
inspiration
memory, nature of
mesmerising, nature of novel
metaphor
moons, shattered
narrative
novel, mesmerising
oblivion, nature of
observations, traveller
personal, nature of reflections
randomness
reading
reflections, personal
rings, of Saturn
Rings of Saturn, The
Saturn
Sebald, W. G.
semi-fiction
shattering, of moons
stream-of-consciousness
survival, nature of
titles, book
tour, walking
travelling
walking

This is an evident absurdity (but it is only a thought experiment). To restore some sense, and reduce the number of entries, we have to reintroduce the notion of relevance. And to do that, we have to have made a judgement of what the book is about. If we know the book is about, say, astronomy, then we might index *rings* and *moons*, because we would expect there to be subentries in due course:

moons
  shattered
  unshattered etc.

If we know the book is about creative writing, we might index *stream-of-consciousness* and *narrative* (among others), for the same reason:

stream-of-consciousness
  in astronomy
  in novels etc.

We know that rings and moons are incidental (of negligible relevance) to a book on creative writing. And vice versa: we know that the notion of narrative is incidental to a book on astronomy.

If we cannot make a judgement of what the book is about, then we cannot easily index it. That is why fiction is so difficult to index: its content cannot be so easily reduced to a single theme, and this makes us pause as we consider what items to select for indexing. And that is why a general reference book, such as the *Penguin Factfinder*, which I edit, is so hard to index. Because it deals with everything, I want to index everything, and that is not possible. Considerations of length, cost and time forbid it – as it is, the index is already 140 pages (about 15 per cent of the book).

Indexing the Web

And this is why it is so difficult to index the World Wide Web in a sensible way. The Web is about everything. And many individual sites and pages of the Web are about everything – in the sense that their content is totally unpredictable. Most blogs are like this. They talk about whatever topic happens to come up, day by day. Social networking sites such as MySpace are like this. Broadcasting sites such as YouTube are like this. But it is not just these personal sites that are multithematic. Most news sites are too, as this selection of headlines from CNN illustrates. First, two-theme:

> Ex-Tiger Fielder says he plans to repay debts (baseball, finance)
> Schwarzenegger backs stem cell plan (politics, medicine)
> Exotic frog invades Georgia (animals, USA)
> Tumor may be linked to cell phone use (phones, medicine)
> Infection risk grows for Hong Kong (medicine, China)

Now three-theme:

> Company blasts ashes into space (space, economics, death)
> Chinese showcase fuel-saving cars (cars, China, energy)
> AirAsia, Malaysian Air discuss cooperation (air travel, Malaysia, politics)

And sometimes even four-theme:

> Student killed during postgame celebration: woman hit by projectile fired by officer; police take full responsibility (baseball, policing, education, safety)

These are examples where the themes are explicit at the outset. Rather more subtle are those where themes are 'buried' in the body of the text. A news item might begin by reporting on a film star's latest movie, but half-way down begin to talk about his impending divorce or his eating habits or whatever. When we take all these possibilities into account, it turns out that it is relatively unusual to find a web page which is strictly monothematic.

There are basically two approaches to indexing 'out there', and neither captures the multithematic character of the Web. One is index maximalism – the Google approach. The software indexes everything apart from a few stop words, such as *the*. We know the strengths and the weaknesses of this. If our query is highly specific, we will get a useful result. Finding Ford Cortinas or Tom Cruise is easy. But if it is not, we will get millions of diverse results, and huge amounts of irrelevance. Finding information on, say, 'main universities in France that teach linguistics' – something I had to do the other day – proved impossible. The more abstract, wide-ranging, ambiguous or metaphorical our enquiry, the more we will end up frustrated. It is not that the pages are not there, it is just that they have not been indexed in a way that anticipates the relevance needs of the user.

The other is index minimalism – an approach found in online advertising, where teams of people scrutinize web pages and make a judgement about what they are about, so that a relevant ad can be placed on the screen. It is an approach that is prone to disaster. For example, a while back I saw a news report on the CNN website about a street stabbing in Chicago. The ads down the side said 'Buy your knives here. Get your knives on eBay.' It is easy to see what had happened. The stupid software had scanned the page, found *knife* as a frequent word, and matched it with the keywords it already had in its ad inventory. Because it did not examine all the words on the page, it was unable to detect that the page was about a homicide, and thus unable to work out that any ads should be about personal safety devices or a career in the police or whatever.

Notice that the maximalist approach cannot solve the minimalist one. If the CNN report has a thousand words, then each of these words could be a trigger for an ad. If it happened to mention that the victim's sweater was covered in blood, then that might generate ads for knitwear. Someone has to go through the report and decide what the report is about and identify which words best capture that aboutness. It has to be a someone. No machine can yet do this. And even humans find it difficult, because there are lots of distracting words in a news report – even on a page which you might think as monothematic, such as a science page.

To illustrate, consider this paragraph, taken from a website on weather:

> Depressions, sometimes called mid-latitude cyclones, are areas of low pressure located between 30° and 60° latitude. Depressions develop when warm air from the sub-tropics meets cold air from the polar regions. There is a favourite meeting place in the mid-Atlantic for cold polar air and warm sub-tropical air. Depressions usually have well defined warm and cold fronts, as the warm air is forced to rise above the cold air. Fronts and depressions have a birth, lifetime and death; and according to the stage at which they are encountered, so does the weather intensity vary.

Which words identify the topic of 'depression'? Some, such as *cyclone*, *warm front* and *cold front*, are clearly highly relevant – they are hardly ever used outside this context. Others, such as *birth*, *lifetime* and *death*, are clearly irrelevant – part of the literary style, but not the topic. And others are of uncertain relevance: *intensity*, *vary*, *areas*, *meeting place*, *mid-Atlantic, cold air*, all of which can be used in several other contexts in the language – *cold air* in relation to air-conditioning, for example, or *mid-Atlantic* in relation to yacht racing. Nor are the terms *front* and *depression* by themselves as helpful as you might think, for they have many other meanings in English. Indeed, type *depression* into Google and you will be swamped with advice about how to cure your mind.

Nonetheless, it ought to be possible to rank the words on a page roughly in order of relevance, with (in this example) *cyclone* towards the top and *the* towards the bottom, and this is what needs to be done if we are to solve the problem of indexing multithematic pages or sites. Indexers are best placed to do this, of course, as indexing is, more than anything else, a matter of judging relevance.

To solve the problem of web indexing, we have to anticipate what people might want to talk about. It sounds like an infinite task, but it isn't, because to talk about anything, people have to use the words of their language, and this is a finite list. Most of the words they need will be found in a medium-sized dictionary (a college dictionary of about 1,500 pages, such as the *Concise Oxford*, contains about 100,000 entries). If we can categorize all the words and senses that are likely to generate search queries, then we have broken the back of the problem. The average number of senses per headword in such a dictionary is 2.4. We are talking of a lexical inventory of about a quarter of a million items, therefore. That is the project I directed in the mid-1990s. It took about five years for a team of lexicographers to go through a dictionary in this way, assigning each word/sense to a taxonomic category (such as weather, botany, psychiatry, or one of its sub-divisions).

The pool of words

Putting this another way, if you want to talk about the weather, what is the pool of words in English from which you must choose? They will be words like *rainy*, *hot, outlook*, *depression*. They will not be words like *bishop*, *army*, *betrayal* and *incognito*. Defining these word-pools, for each topic, in a taxonomy, is the nature of the task. Taxonomies can go on for ever, but one starts at the top and works down. The taxonomy I use currently has some 2,500 categories. Each category has a word-pool of between 100 and 200 items, using both proper names and English lexemes.

It is a never-ending project, of course, because language changes and the world changes, and what is a relevant word for a category in one year might not be so for the next. This especially happens in political categories, where presidents and prime ministers change, and old names on web pages are replaced by new ones. But all categories need monitoring, especially in the commercial world, where new brands, models and product names are routine. To take an example: our word-pool for 'weapons' in 2000 did not have *weapons of mass destruction*. It does now.

All this has to be done by humans at present. There are ways in which we can teach computers to replicate what humans do, but for this task, not yet. Most software programs still use simple algorithms, such as looking for the rarest words or the most frequent words. Neither suffices. Simple logic never works, because language follows principles that are alien to the way computers operate: stylistic principles, in particular. In a web page reporting a football match, for instance, you might expect the word *football* to turn up a lot. It does not. Because everyone knows the page is about football, it is hardly ever mentioned. Likewise, although football is all about kicking a ball, a verb like *kick* is not common on a football page. When a footballer kicks a ball, the report says he shoots, lifts, slams, hammers the ball – never boringly kicks it.

A linguistic perspective, informed by sociolinguistic and stylistic considerations, is crucial in web indexing. That is the only way to avoid the crass errors that machines routinely make. For example, without a good linguistic awareness, the computer (i.e. the people who program the computer) will assume that such word-groups as *operation*, *operate* and *operator* are all closely linked, so that when you find one you will find the others. But it is not like that. Surgeons perform operations and they operate, but they are not operators. And telephone operators do not carry out operations.

A set of relevance judgements has to be tested, of course. This is how we do it, in the approach I have been developing over the past few years. We choose a topic (such as weather), and, using our dictionaries, linguistic intuitions, general knowledge and a sample of websites, identify the words (technically, lexemes and proper names) that we consider to be the most relevant. We give them a weighting, reflecting our sense of just how relevant they are to the topic. We then collect these items in a file and use this file as a filter. We devise software that applies this filter to web pages. If we have got our word-pool right, then it should correctly identify any meteorological web page as being about the weather, and ignore any web page that is not. If it misclassifies, we have to look at the page to see why, and alter our word-pool so that it works better next time.

We have so far tested about 2,500 categories in this way, covering the range of subject-matter found in our general encyclopedias (the Cambridge and Penguin families). An important point to appreciate is that every web page is tested against all 2,500 categories. This is the only way to capture the multithematic character of the Web. The web page above about the student who was killed actually contains four themes, and if its content is to be captured accurately, the results need to show four relevant classifications – in this case, baseball, policing, education and safety. The software behind this needs to be good, of course, to ensure rapid and scaleable results (as stressed by Richard Northedge in the last *Indexer*). It can classify a web page in this way in a tenth of a second.

Applications

This kind of linguistic indexing has all kinds of applications, apart from improving results in search engine enquiries and contextual advertising. It can be used for automatic document classification – identifying which groups of electronic documents go together, with respect to a particular theme. There is also a forensic set of applications: for example, it is possible to monitor the lexical content of conversations in real time to determine whether they contain sensitive information, as in paedophile intrusion into child chatrooms. And it is also very easy to identify words that are felt to be objectionable, so that sites that contain them are avoided in an enquiry. This is especially important in the advertising world, where advertisers might not want their products to appear alongside certain kinds of content, such as adult sites or sites expressing extreme political views.

We are at the beginning of a long road. A total of 2,500 categories is tiny, compared with the number of discriminations yet to be made, as we follow taxonomies down to lower levels. If we look at the number of categories found in a 'bottom-up' system such as Dmoz, we are talking about tens of thousands. And it all has to be done for different languages, for the Web is a hugely multilingual world. It is exciting to have been in at the start of this world, and especially rewarding to know that the progress I have made could not have been achieved without my experience as an indexer.

Adapted from a paper given at the SI 50th Anniversary Conference, 13 July 2007.

***David Crystal** is honorary professor of linguistics at the University of Bangor, and the author (2008) of* Think on my words: exploring Shapespeare's language *and* Txtng: the Gr8 Db8.
Email: davidcrystal@googlemail.com