We have thus far viewed a document as a sequence of terms. In fact, most

documents have additional structure. Digital documents generally encode,

METADATA in machine-recognizable form, certain metadata associated with each docu-

ment. By metadata, we mean specific forms of data about a document, such

as its author(s), title and date of publication. This metadata would generally

FIELD include fields such as the date of creation and the format of the document, as

well the author and possibly the title of the document. The possible values

of a field should be thought of as finite – for instance, the set of all dates of

authorship.

Consider queries of the form “find documents authored by William Shake-

speare in 1601, containing the phrase alas poor Yorick”. Query processing then

consists as usual of postings intersections, except that we may merge post-

PARAMETRIC INDEX ings from standard inverted as well as parametric indexes. There is one para-

metric index for each field (say, date of creation); it allows us to select only

the documents matching a date specified in the query. Figure 6.1 illustrates

the user’s view of such a parametric search. Some of the fields may assume

ordered values, such as dates; in the example query above, the year 1601 is

one such field value. The search engine may support querying ranges on

such ordered values; to this end, a structure like a B-tree may be used for the

field’s dictionary.

ZONE Zones are similar to fields, except the contents of a zone can be arbitrary

free text. Whereas a field may take on a relatively small set of values, a zone

can be thought of as an arbitrary, unbounded amount of text. For instance,

document titles and abstracts are generally treated as zones. We may build a

separate inverted index for each zone of a document, to support queries such

as “find documents with merchant in the title and william in the author list and

the phrase gentle rain in the body”. This has the effect of building an index

that looks like Figure 6.2. Whereas the dictionary for a parametric index

comes from a fixed vocabulary (the set of languages, or the set of dates), the

dictionary for a zone index must structure whatever vocabulary stems from

the text of that zone.

In fact, we can reduce the size of the dictionary by encoding the zone in

which a term occurs in the postings. In Figure 6.3 for instance, we show how

occurrences of william in the title and author zones of various documents are

encoded. Such an encoding is useful when the size of the dictionary is a

concern (because we require the dictionary to fit in main memory). But there

is another important reason why the encoding of Figure 6.3 is useful: the

WEIGHTED ZONE efficient computation of scores using a technique we will call weighted zone

SCORING scoring.