Range Results in XML Retrieval

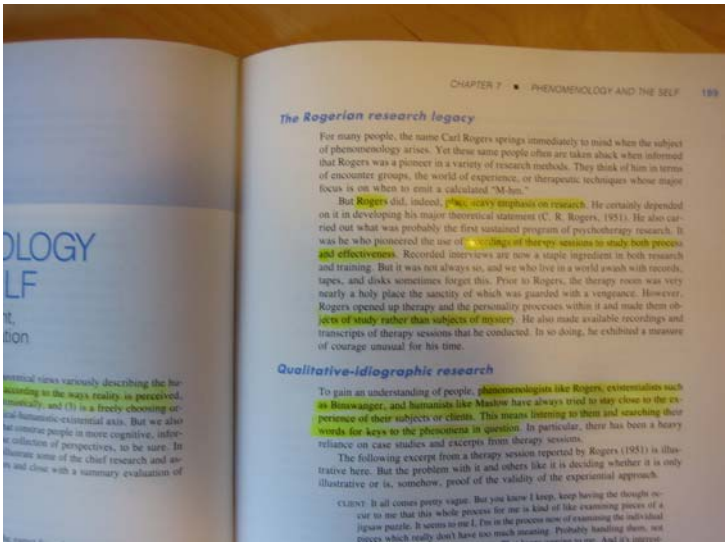
Charles L. A. Clarke
School of Computer Science
University of Waterloo
Waterloo, Canada

Ranges

- “pages 42-57”
- “chapters 2 and 3”
- “sections 4.2 to 4.7”
- “the first two sentences of the third paragraph”

all expressible in XPath

Highlighting



CHAPTER 7 ■ PHENOMENOLOGY AND THE SELF 199

The Rogerian research legacy

For many people, the name Carl Rogers springs immediately to mind when the subject of phenomenology arises. Yet these same people often are taken aback when informed that Rogers was a pioneer in a variety of research methods. They think of him in terms of encounter groups, the world of experience, or therapeutic techniques whose major focus is on when to emit a calculated "M-hm."

But Rogers did, indeed, place heavy emphasis on research. He certainly depended on it in developing his major theoretical statement (C. R. Rogers, 1951). He also carried out what was probably the first sustained program of psychotherapy research. It was he who pioneered the use of recordings of therapy sessions to study both process and effectiveness. Recorded interviews are now a staple ingredient in both research and training. But it was not always so, and we who live in a world awash with records, tapes, and disks sometimes forget this. Prior to Rogers, the therapy room was very nearly a holy place the sanctity of which was guarded with a vengeance. However, Rogers opened up therapy and the personality processes within it and made them objects of study rather than subjects of mystery. He also made available recordings and transcripts of therapy sessions that he conducted. In so doing, he exhibited a measure of courage unusual for his time.

Qualitative-idiographic research

To gain an understanding of people, phenomenologists like Rogers, existentialists such as Binswanger, and humanists like Maslow have always tried to stay close to the experience of their subjects or clients. This means listening to them and searching their words for keys to the phenomena in question. In particular, there has been a heavy reliance on case studies and excerpts from therapy sessions.

The following excerpt from a therapy session reported by Rogers (1951) is illustrative here. But the problem with it and others like it is deciding whether it is only illustrative or is, somehow, proof of the validity of the experiential approach.

CLIENT: It all comes pretty vague. But you know I keep, keep having the thought occur to me that this whole process for me is kind of like examining pieces of a jigsaw puzzle. It seems to me I, I'm in the process now of examining the individual pieces which really don't have too much meaning. Probably handling them, not

NOT arbitrary passage retrieval

Still want to present the user with *coherent* document components:

sections, pages, subsections,
paragraphs, chapters, articles…

Range Results at INEX

INEX 2004 adhoc relevance judgments:

Of the 5229 elements judged as highly exhautive and specific, at least 1700 (32%) are part of a larger range of elements with identical tag names.

Example: Topic 165

(3,3) elements of ex/2001/x3047.xml:

```
/article[1]/bdy[1]/sec[5]/ss1[1]/p[3]
/article[1]/bdy[1]/sec[5]/ss1[1]/p[4]
/article[1]/bdy[1]/sec[5]/ss1[1]/p[5]
/article[1]/bdy[1]/sec[5]/ss1[1]/p[6]
```

Might be better expressed as:

```
/article[1]/bdy[1]/sec[5]/ss1[1]/p[3 to 6]
```

Range Results in XPath

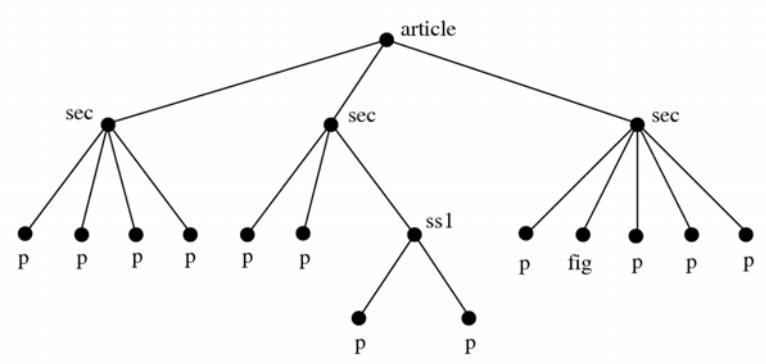
XPath may be too expressive:

```
//sec/p[2]
```

The second paragraph of every section has limited value as a retrieval result.

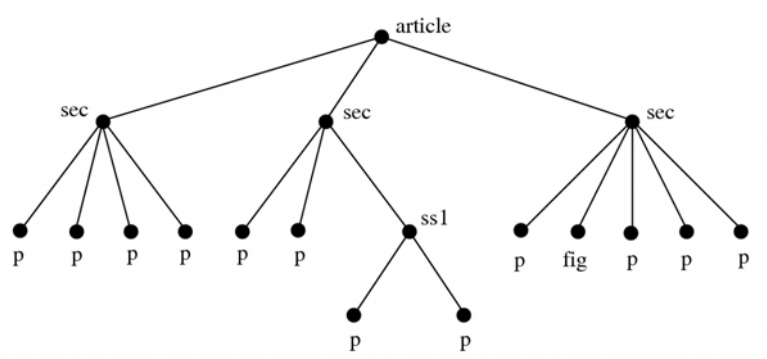
but expressing range results is difficult

First three paragraphs of 1st section



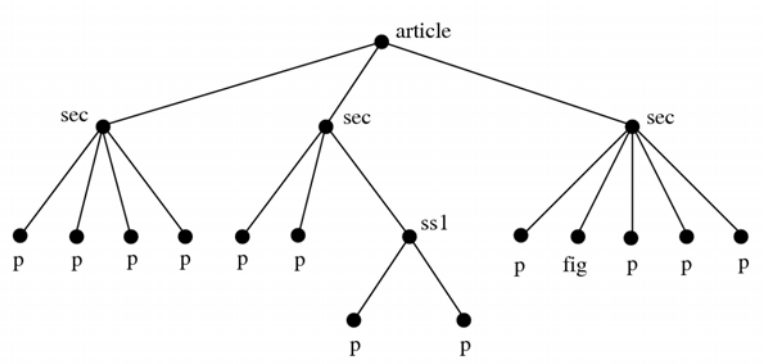

```
/article[1]/section[1]/p[position() <= 3]
```

First three paragraphs of 2nd section



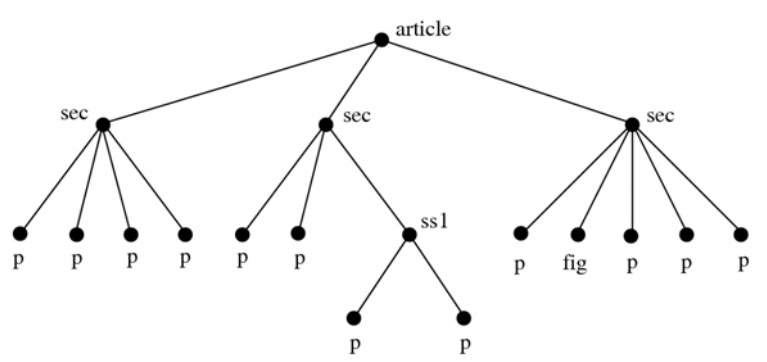

```
/article[1]/section[2]/p[position() <= 3]
```

First three paragraphs of 2nd section




```
/article[1]/section[2]//p[position() <= 3]
```

First three paragraphs of 3rd section




```
/article[1]/section[3]//p[position() <= 3]
```

Specifying ranges by endpoints

Given two location paths, *X* and *Y,* the XPath 2.0 expression

X`/following::*[.<<` *Y*`]`

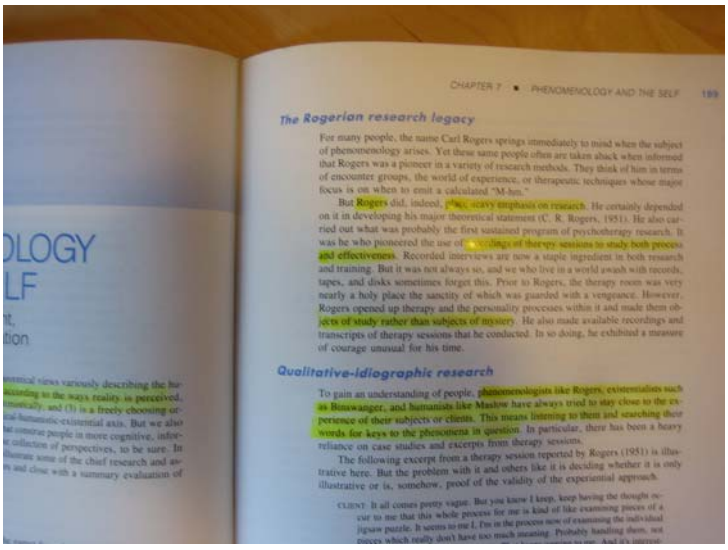
includes all the elements between them.

Results may be specified by (X,Y) pairs.

Evaluation

- $O(N^2)$ range components
- Reusing the test collection?
- Extending INEX methodology?

Evaluation



CHAPTER 7 ■ PHENOMENOLOGY AND THE SELF 199

The Rogerian research legacy

For many people, the name Carl Rogers springs immediately to mind when the subject of phenomenology arises. Yet these same people often are taken aback when informed that Rogers was a pioneer in a variety of research methods. They think of him in terms of encounter groups, the world of experience, or therapeutic techniques whose major focus is on when to emit a calculated "M-hm."

But Rogers did, indeed, place heavy emphasis on research. He certainly depended on it in developing his major theoretical statement (C. R. Rogers, 1951). He also carried out what was probably the first sustained program of psychotherapy research. It was he who pioneered the use of recordings of therapy sessions to study both process and effectiveness. Recorded interviews are now a staple ingredient in both research and training. But it was not always so, and we who live in a world awash with records, tapes, and disks sometimes forget this. Prior to Rogers, the therapy room was very nearly a holy place the sanctity of which was guarded with a vengeance. However, Rogers opened up therapy and the personality processes within it and made them objects of study rather than subjects of mystery. He also made available recordings and transcripts of therapy sessions that he conducted. In so doing, he exhibited a measure of courage unusual for his time.

Qualitative-idiographic research

To gain an understanding of people, phenomenologists like Rogers, existentialists such as Binswanger, and humanists like Maslow have always tried to stay close to the experience of their subjects or clients. This means listening to them and searching their words for keys to the phenomena in question. In particular, there has been a heavy reliance on case studies and excerpts from therapy sessions.

The following excerpt from a therapy session reported by Rogers (1951) is illustrative here. But the problem with it and others like it is deciding whether it is only illustrative or is, somehow, proof of the validity of the experiential approach.

CLIENT: It all comes pretty vague. But you know I keep, keep having the thought occur to me that this whole process for me is kind of like examining pieces of a jigsaw puzzle. It seems to me I, I'm in the process now of examining the individual pieces which really don't have too much meaning. Probably handling them, not

Evaluation

Given a range component *X* with relevant text highlighted:

$$\frac{\text{size of relevant text in } X}{\text{size of } X}$$

Specificity?

Evaluation

Given a range component *X*, contained in document *D*, with relevant text highlighted:

$$\frac{\text{size of relevant text in } X}{\text{size of relevant text in } D}$$

Exhaustivity? Coherence?

Questions?

- Would range results be useful?
- What other evaluation measures could be used?
- What retrieval techniques could be used?

Discussion

Range Results in XML Retrieval

Charles L. A. Clarke
School of Computer Science
University of Waterloo
Waterloo, Canada