

Range Algebra

Gavin Thomas Nicol
CTO
Red Bridge Interactive, Inc.

Agenda

Presentation Agenda

Motivation

Range Algebra

- Core Range Algebra

- Range Constructors

- Attributed Range Algebra

Status & Future Work

Motivation

Which came first, text or markup?

A text is some form of written communication

Markup is added to the text to

- Correct, or annotate the text

- Add semantics to the text.

- Aid in formatting.

Why is it that formal models of documents model the markup?

```
\begin{section}
```

```
\title{Introduction}
```

With the advent of the WWW, hypermedia has been getting a great deal of attention in the media and in the workplace as a way for managing and navigating information. Despite the fact that there has been a very active research community for over twenty years, hypermedia is seen as a new technology, and is often misrepresented.

This article is an attempt to clarify the notion of exactly what hypermedia is, to give an overview of the vast amount of research done to date, to review the various implementation strategies and technologies, to discuss some of the design issues, and to provide some glimpses into the future of hypermedia.

```
\end{section}
```

Is a document *really* a tree?

Simplistically, yes.
Practically, no
(DeRose/Renear et al).
Documents have

Annotations

Overlapping structures

Cross references

Etc.

Why is it that most document
models are trees?

Doesn't that complicate
things?

```
<section>
<title>Introduction</title>
<p>With the advent of the WWW,
<i>hyper</i>media has been getting a great
deal of attention in the media and in the
workplace as a way for managing and
navigating information. Despite the fact that
there has been a very active research
community for over twenty years, hypermedia
is seen as a new technology, and is often
misrepresented. </p>

<p>This article is an attempt to clarify the
notion of exactly what hypermedia is, to give
an overview of the vast amount of research
done to date, to review the various
implementation strategies and technologies,
to discuss some of the design issues, and to
provide some glimpses into the future of
hypermedia. </p>
</section>
```

Is there a canonical document structure?

Simplistically, yes.

Practically, no

Semantic interpretation is in the eye of the beholder.

Times change, things evolve, new fashions emerge.

Why do most markup languages force single type conformance?

Do we really need pointy brackets?

```
<section>
<title>Introduction</title>
<p>With the advent of the WWW,
<i>hyper</i>media has been getting a great
deal of attention in the media and in the
workplace as a way for managing and
navigating information. Despite the fact that
there has been a very active research
community for over twenty years, hypermedia
is seen as a new technology, and is often
misrepresented. </p>

<p>This article is an attempt to clarify the
notion of exactly what hypermedia is, to give
an overview of the vast amount of research
done to date, to review the various
implementation strategies and technologies,
to discuss some of the design issues, and to
provide some glimpses into the future of
hypermedia. </p>
</section>
```

Summary

Most current models of documents are

- Limiting & inaccurate

- Fail to deal with *text* well (data centric)

Documents are *not* always inherently trees

- Though it is often very convenient to look at them as such.

Documents do *not* necessarily have a canonical format.

- Implied semantics are as important as explicit semantics.

Range Algebra

Range Algebra Is

An alternate formal model of data

- Not limited to trees, or text

- Able to handle both explicit and implicit structures

- A basis for validation of partially marked-up texts

- A basis for formal integration of different markup languages

- Displays natural closure in operations

A different way of thinking about data

- Markup is ephemeral

- Emphasis on type by projection

- Emphasis on graceful evolution of data structures

Data format independent

As old as the hills...

Core Range Algebra

Sequence = completely ordered finite set of items from an alphabet

Range = tuple{ start, length } (access as R.start, R.length)

$\{0,3\}$



0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
T	H	E		Q	U	I	C	K		B	R	O	W	N		F	O	X



Sequence of characters (file/string)

Core Range Algebra (contd.)

Sequence operations

Difference, Intersection, Union, Concatenation
etc.

Range operations

StartsBefore, StartsAfter, StartsWithin
EndsBefore, EndsAfter, EndsWithin, Within
Move, Resize, Extract, Normalize
Etc.

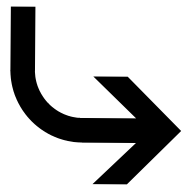
Other operations

StartOrder, EndOrder,
Ancestor, Descendant
Parent, Child

Range Constructors

Range Constructor = A function that, given a sequence, returns a sequence of ranges

0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
T	H	E		Q	U	I	C	K		B	R	O	W	N		F	O	X



{0,3}	{4,5}	{10,5}	{16,3}
-------	-------	--------	--------

Regular Expressions

Regular expressions are defined in terms of an alphabet (a finite set of letters) as is the sequence data type.

Regular expressions are the basis for a powerful range constructor.

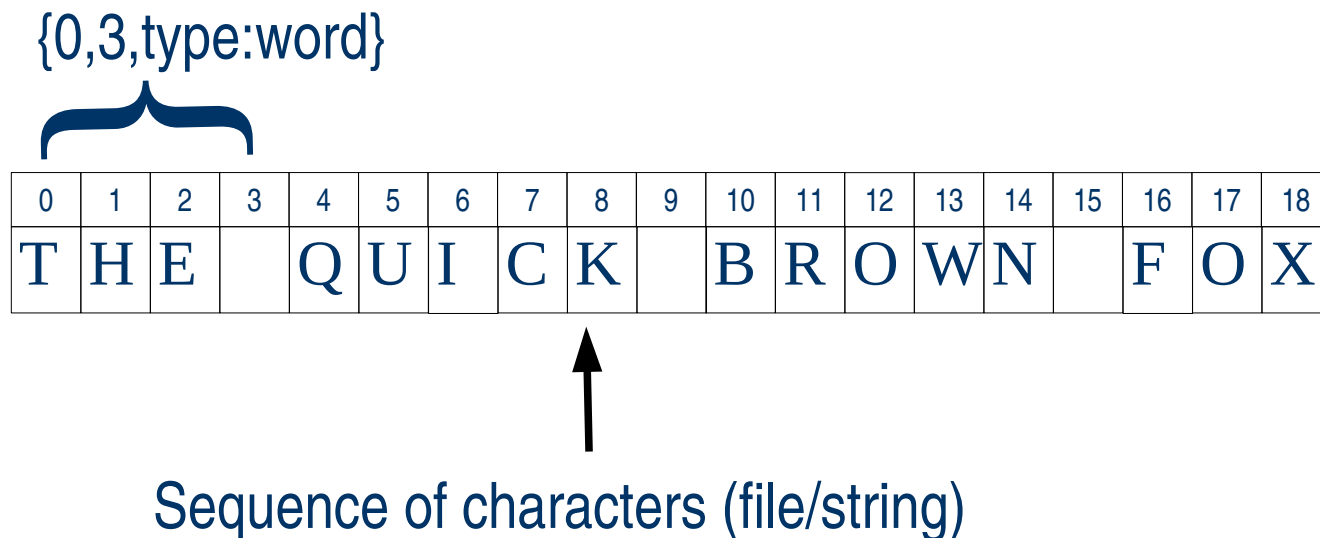
```
[^a-zA-Z][a-zA-Z]+[^a-zA-Z]  
Construct($1.start, $3.end)
```

```
Union(Construct(S, "*day"),  
      Union(Construct(S, "(Tues|Fri)day"),  
            Construct(S, "(Mon|Wednes)day")))
```

Attributed Range Algebra

Extends the basic range type to allow ranges to have an arbitrary set of attributes (access as *R.name*)

Provides the basis for semantic interpretation of ranges and validation/manipulation of structure



Attributed Range Algebra (contd.)

All Core Range Algebra operations are applicable to Attributed Range Algebra, including range constructors.

Use of regular expressions is purely a syntax issue

Proposal: extend POSIX regular expressions to allow embedded feature logic expressions.

```
[^a-zA-Z][a-zA-Z]+[^a-zA-Z]  
    Construct($1.start, $3.end,  
              Attribute("type", "word"))  
  
[%type=word%]+[%type=period%]  
    Construct($1.start, $2.end,  
              Attribute("type", "sentence"))
```


Range Algebra & XML

Possible to parse XML using pipelined regular expressions and Attributed Ranges (extension of Shallow Parsing)

Range Algebra provides a natural model for IR operations

Any well-formed XML (core) document can be expressed in Range Algebra in a structure that mimics the node structure

Much of XQuery could be defined in terms of Range Algebra

Other structured/semi-structured data formats can also be expressed as Ranges corresponding to nodes.

Formal basis for incremental/parallel transformations to/from XML

Formal basis for external markup and regular fragmentations

Range Algebra provides a natural model for XLink ranges.

Status and Future Work

Status & Future Work

Status

- Core Range Algebra and Attributed Range Algebra models available in draft form

- Prototyped regular expression range constructors and XML range constructor

- Currently implementing tool chain based on Range Algebra

Future work

- Complete formal definitions

- Proof that RA can be used to express other formal models

- Formalise type by projection model

- Make proof-of-concept tools available