

Sebastian Maneth, Kim Nguyen / NICTA, Neville Roach Lab, Kensington

Software for **super-fast searches of large, complex XML** data supporting applications for **life sciences, publishing, E-commerce, market analytics, financial derivative analysis** and other fields. – Get *more* from your data!



What's Under the Hood

- Databases (e.g. Oracle) are good for tabled search
- Search engines (e.g. Google) are good for keyword search
- The exploding “middle world” of **XML based data needs both!**
- Our engine: revolutionary design from ground up
- Super-fast tree and text indexes
- Unique query compiler – patent pending – leverages the power of the indexes
- Based on path-breaking “automata” research

At a Glance

- Capable of rapid search over **highly structured** data such as tax forms, product catalogs, linguistic data, genomic data, etc.
- High speed, complex searches
- Allows sophisticated search for products (“Intelligent E-Bay”)
- Can replace expensive Database & Search solutions

Looking Ahead

What we have:

- Ultra-fast large-scale XML search
- Operates in main memory providing near real-time capability
- Usable for **E-commerce** and mobilizing semantic web applications

What we want:

- Work with market leaders on killer applications
- Engage with industry with a view to rapid commercialization



Super-fast search over large XML data.

What is difficult about XML?

(1) Represents **tree shaped** data (possibly deep hierarchical structure)

- Trees do not fit well into tables of conventional databases!
- Need to implement a succinct tree storage from scratch

Tree Storage

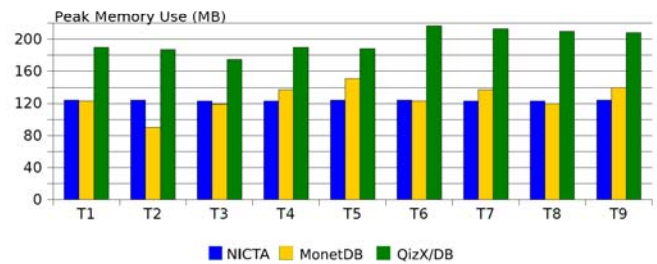
- currently uses Sadakane's parenthesis structures.
- can ultimately be replaced by homegrown (NICTA-patented) tree compression technology
- new challenges (data structure): want to "jump" to nodes in the compressed tree (shortcut the moving along edges). Need succinct index to keep "jump information".

Text Storage

(2) Represents **text data** (usually >>50% of the XML).

- Use state-of-the-art text indexes

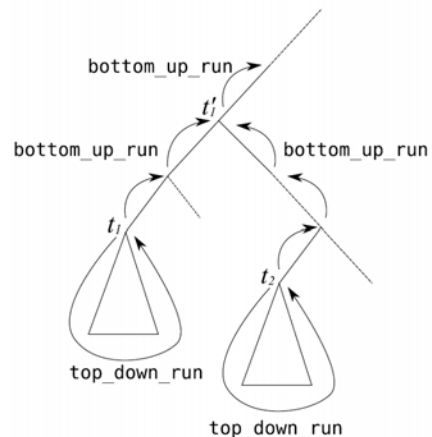
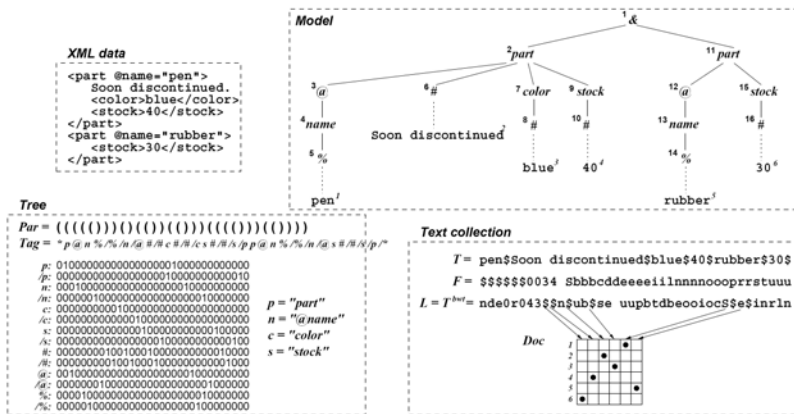
	T1	T2	T3	T4	T5	T6	T7	T8	T9
Text query	10	4	4	0.4	0.4	9	6	59	0.7
Auto. run	28	8	3	0.9	1.5	17	108	96	2
NICTA: Total	38	12	7	1.3	1.9	26	144	175	2.7
MonetDB	336	118	117	252	301	180	256	473	505
Qizx/DB	108	10	6	99	107	244	259	2469	1397
# of results	1493	438	438	32	32	680	6935	6685	36



Query Intelligence

(3) Comes with **new query languages** (XPath and XQuery)

- Build new **XPath engine**; compile to operations of the indexes.
- Engine switches dynamically between indexes, based on unique "counting statistics" of the XML



From imagination to impact