Statistical Analysis of Real XML Data Collections

Irena Mlynkova
Kamil Toman
Jaroslav Pokorny

Department of Software Engineering
Faculty of Mathematics and Physics
Charles University

Structure of The Talk

- Introduction
- Crawled XML web
- Real XML documents
  - General analyses,
  - Recursion
  - Relational patterns
  - Mixed content
- XML statistics utilization
- Conclusion

Introduction

- XML and related technologies - a leading role among standards for data representation
- Semistructured, self-descriptive
- Possibility to express the allowed structures
  - DTD, XML Schema, Relax NG, ...
- Different techniques are needed for
  - managing, processing, querying, updating
  - compressing, versioning
- ...
- XML analyses can be used for
  - XML pattern processing
  - Repository adaptation
  - Benchmarking
  - ...

General Processing Techniques

- “As general as possible”
  - correct at first glance
  - unnecessarily complex
  - often inefficient
- With restricted features
  - more down-to-earth
  - more effective
  - restrictions are often “unnatural” (based on particular technique)
  - effectiveness suffers when data do not correspond to expectations

Web XML Document Analysis

- L. Mignet, D. Barbosa, P. Veltri:
  The XML Web: The First Study (ACM 2003)
  - document size varies from 10B to 4.6kB
  - for documents up to 4kB the number of element nodes is about 50%, the number of attributes about 30%
    - 18% of elements have no attributes
  - mixed content found in 72% of documents (5% of contents)
  - 99% of documents shallow (depth < 8)
    - average depth 4
  - only 260 total different recursive elements found
  - in 98% of recursive documents there is only one recursive element
  - 95% of recursive documents do not refer DTD or XSD

Real XML Documents

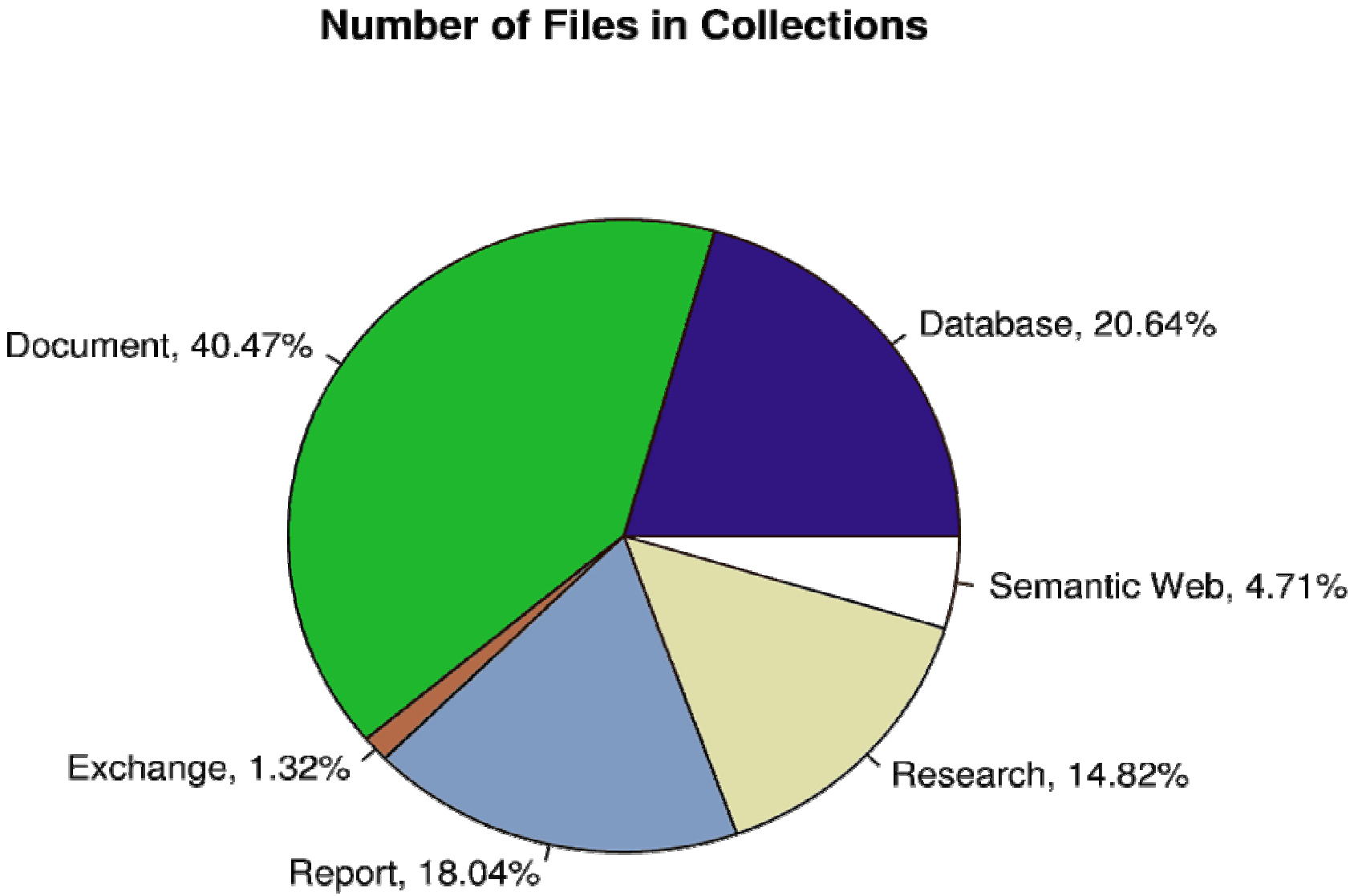
- Classification
  - *data-centric documents (dat)*
    - *database exports, IMDb, list of employees, ...*
  - *document-centric documents (doc)*
    - *Shakespeare's plays, XHTML documents, novels, docbook, ...*
  - *data exchange documents (ex)*
    - *medical information, exchange formats, ...*
  - *reports (rep)*
    - *overviews or summaries*
  - *research documents (res)*
    - *docs with special structures, DNA/RNA, NASA findings, ...*
  - *semantic web documents (sem)*
    - *RDF, OWL, DAML, ...*

Real XML Documents

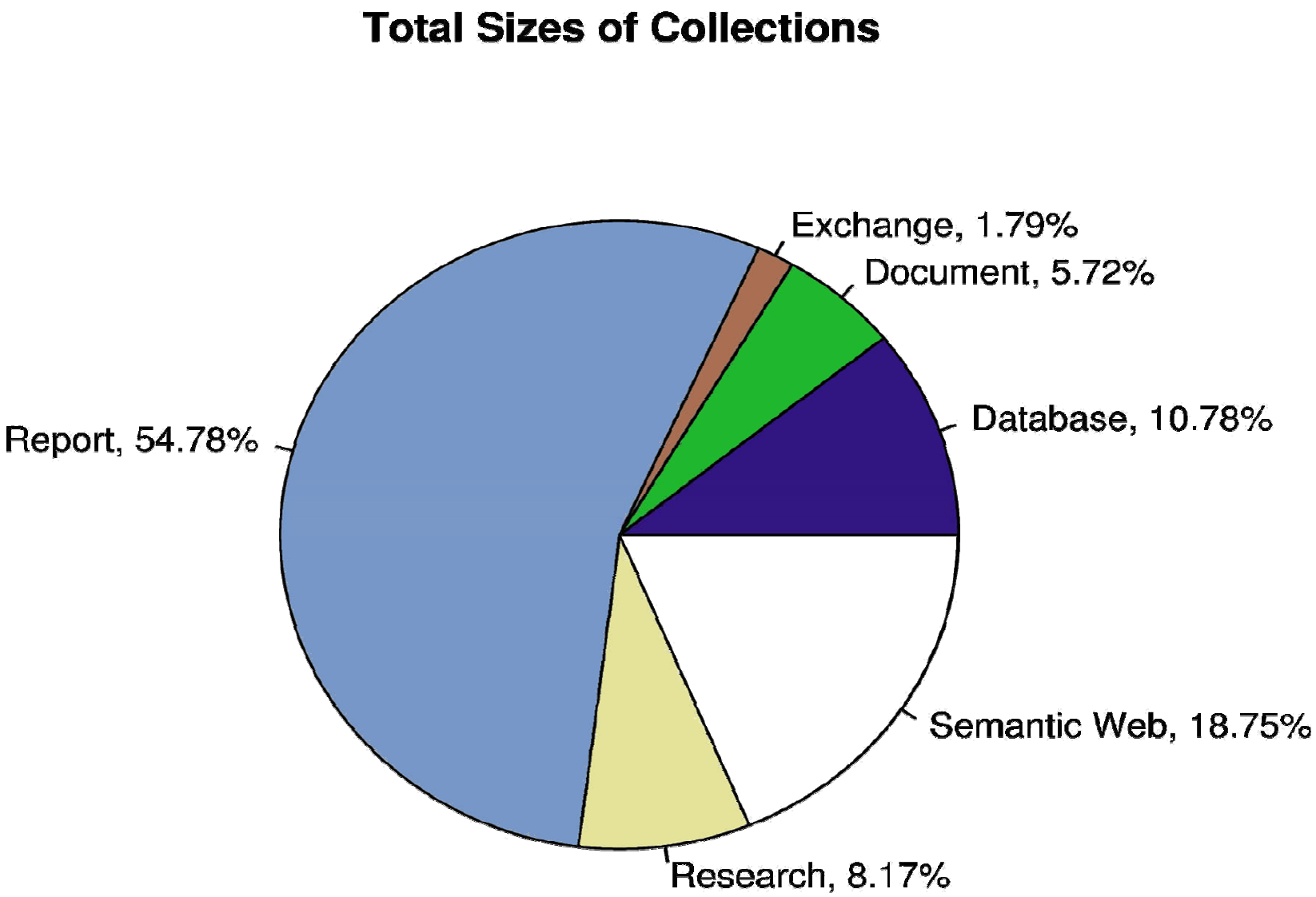
Statistics		Results
Number	Number of XML documents	16,534
	Number of XML collections	133
Size	Total size of documents (MB)	20,756
	Minimum size of a document (B)	61
	Maximum size of a document (MB)	1,971
	Average size of a document (MB)	1.3
	Median size of a document (kB)	10
Schema	Documents with DTD (%)	74.6
	Documents with XSD (%)	38.2
	Documents without DTD/XSD (%)	7.4

General statistics for XML data

Real XML Documents



Real XML Documents



Real XML Documents

Statistics		dat	doc	ex	rep	res	sem
Max. number of elements		402	4,085	37,502	309,379	427	112,942
Max. number of attributes		9	1,675	5,182	37,815	129	37,996
Max. number of empty elements		3	361	123	16,348	6	23,635
Max. number of mixed elements		0	302	21	0	1	0
Max. number of distinct el. names		81	48	58	388	44	144
Max. number of rec. elements		0	3	2	0	0	0
Max. number of distinct paths		79	96	67	312	30	143
Depth of document	Avg.	5	7	5	5	5	5
	Max.	5	13	9	6	7	6

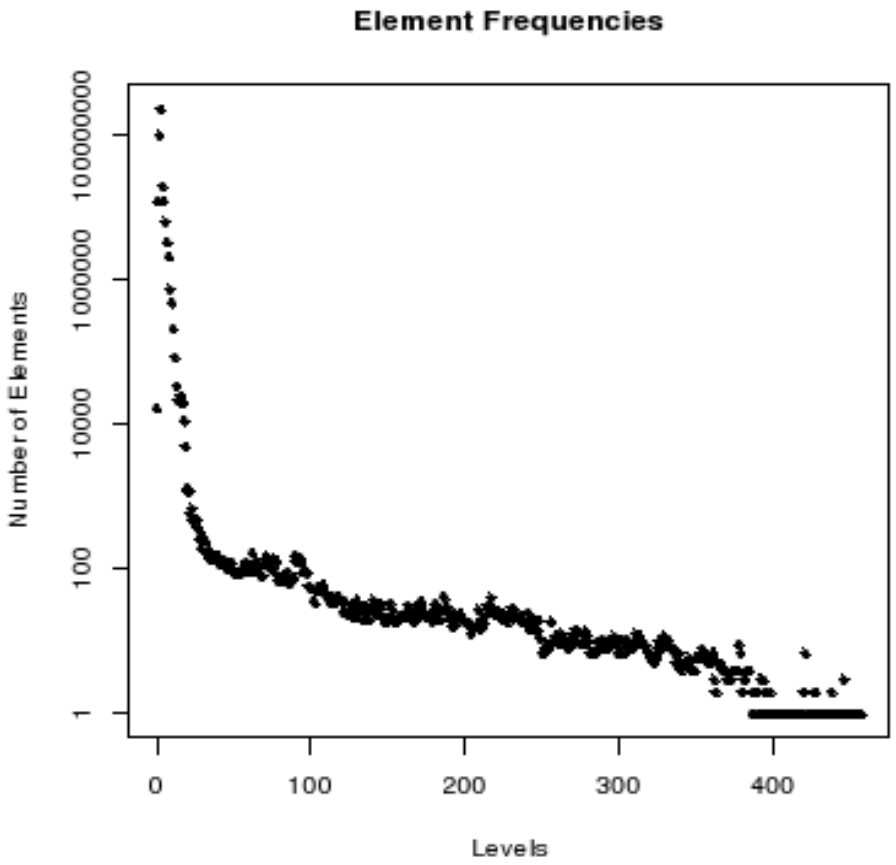
Global statistics for 95% XML documents

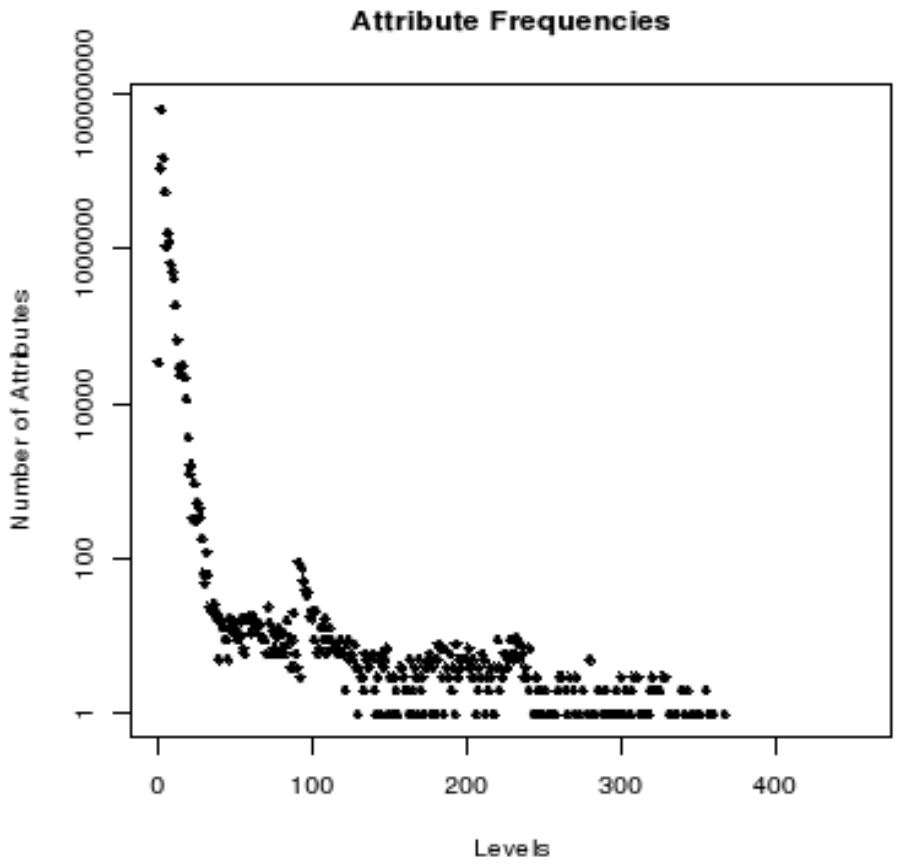
Real XML Documents

	Statistics	dat	doc	ex	rep	res	sem
Doc.	Num. of elements	23,132,565	267,632	2,911,059	1,957,637	21,305,818	25,548,388
	Num. of attributes	33,660,779	102,945	857,691	208,265	2,189,859	10,228,483
	Distinct elem. names	81	134	146	461	210	1,410
	Num. of distinct paths	434	2,086	144	373	426	2,534
	Depth of document	12	459	14	6	19	11
Sch.	Distinct elem. names	76	377	523	3,213	250	-
	Num. of distinct paths	115	11,994	1,665	3,137	568	-
	Depth of schema	12	81	79	5	15	-

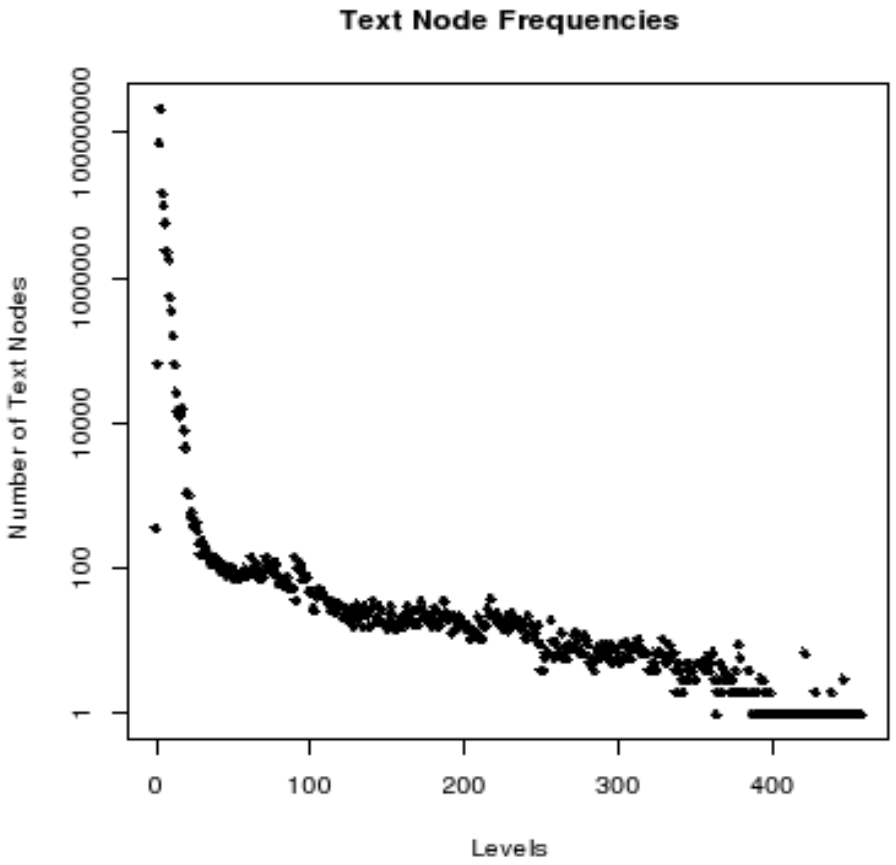
Maximum values of global statistics

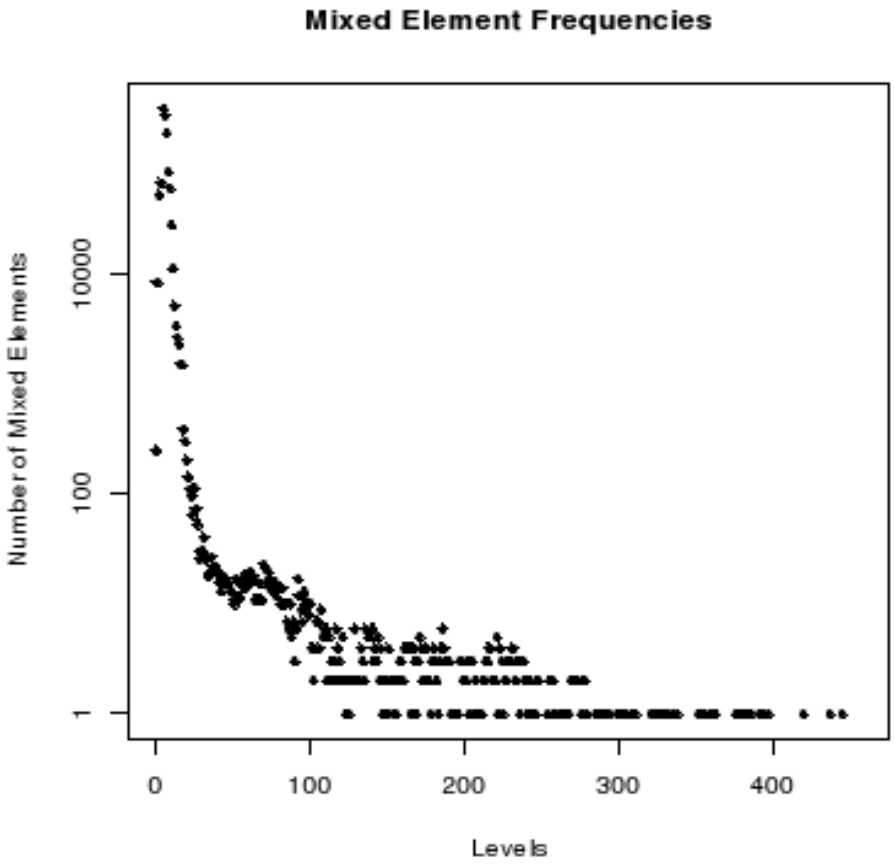
Real XML Documents



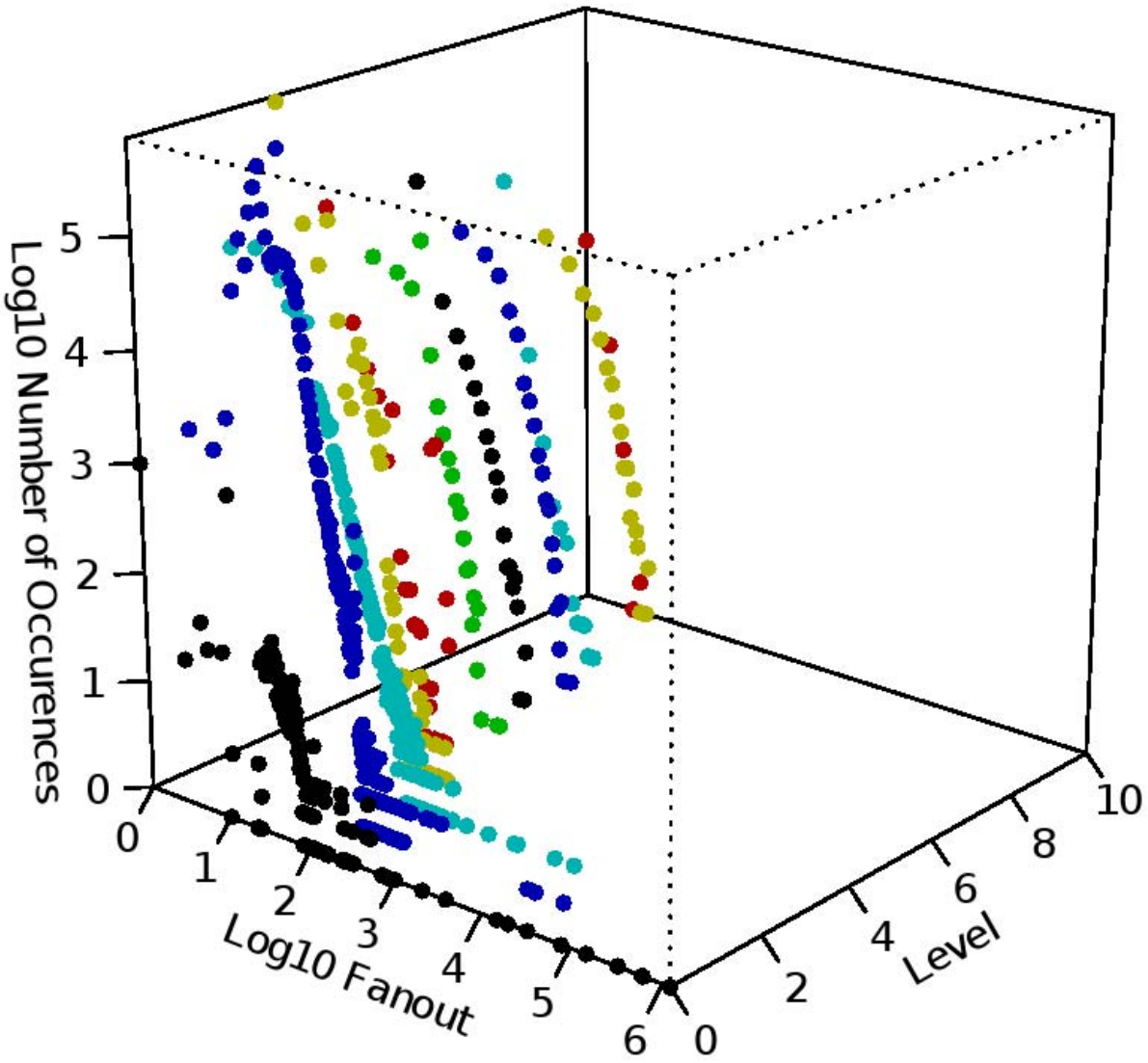


Real XML Documents





Real XML Documents – Fanout Distribution (Database)



Real XML Documents

- New constructs
  - *trivial element* – content model *a := e | pcdata*
  - *simple element – consists only of trivial elements*
  - *complex elements – otherwise*
- *Recursivity*
  - *trivial - “selfrecursive”, no branching*
    - *<a><a><a>...*
  - *linear – similar to trivial but can intermix with regular elements, single recursive element*
    - *<a><a>...<c/>*
  - *pure – single recursive element, branching possible*
    - *<a><a>...<a>...<c/><a>...<d/>*
  - *general – more than one recursive element*

Real XML Documents

		dat	doc	ex	rep	res	sem
Doc.	T	0.06	2.38	3.67	-	0	0.27
	L	0.06	19.92	32.57	-	0.65	2.52
	P	0.03	18.76	22.48	-	0	1.46
	G	0.06	16.20	7.80	-	0.04	0
Sch.	T	0	0	0	-	0	-
	L	0	0	0	-	14.29	-
	P	0	2.94	7.89	-	28.57	-
	G	12.50	85.29	13.16	-	28.57	-

Exploitation rate of types of recursions (%)

		dat	doc	ex	rep	res	sem
Doc.	T	0.2	5.0	6.4	-	0	1.0
	L	0.5	65.3	45.7	-	66.7	92.6
	P	0.7	12.7	26.9	-	0	6.4
	G	98.5	17.0	21.0	-	33.3	0
Sch.	T	0	0	0	-	0	-
	L	0	0	0	-	2.9	-
	P	0	0.1	1.0	-	20.6	-
	G	100.0	99.9	99.0	-	76.5	-

Percentage representation of types of recursion (%)

Real XML Documents

- *Shallow Relational Patterns*
  - <a>
    one <!-- trivial elements -->
    two
    three

- *Relational Patterns*
  - <x>
    <a>xxx <!-- trivial elements -->
    yyy <!-- no repetition -->
    <c>zzz</c>
    </x>
    <x>
    <a>111 <!-- trivial elements -->
    <c>333</c> <!-- missing elements allowed -->
    </x>

Real XML Documents

Statistics		dat	doc	ex	rep	res	sem
Elements involved		29.23%	6.23%	29.53%	94.29%	22.66%	41.56%
Number of occurences		170,744	154,133	185,358	40,276	619,272	716,038
Repetition	Avg.	10.5	3.3	5.8	322.7	5.1	8.8
	Max.	600,572	1,254	615	102,601	15,814	16,500
Fan-out	Avg.	3.6	1.5	2.2	6.2	2.3	3.5
	Max.	33	10	18	26	51	113

Relational pattern statistics for XML documents per category

Real XML Documents

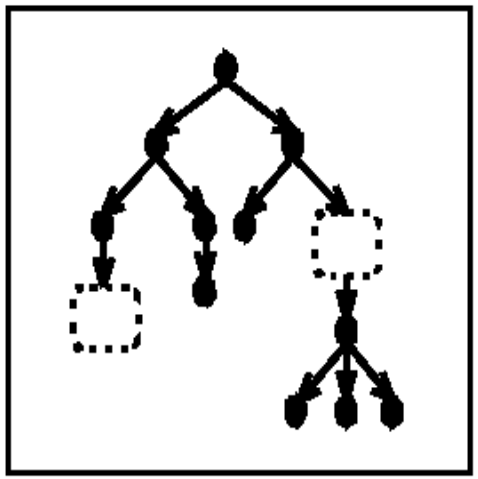
- *Mixed elements*
  - <text><par>
    Some semistructured text including special formatting
    <table><tr><td></td>...</tr>...</table>
    and other complex stuff
    </par><par> ... </par> ...
    </text>
- *Simple mixed elements*
  - <text>Hello bold world!</text>

Statistics		dat	doc	ex	rep	res	sem
Depth	Avg.	1.8	4.1	1.0	-	1.9	1.2
	Max.	6	448	5	-	2	3
Simple mixed contents (%)		55.9	79.4	99.6	-	1.9	78.4

Mixed-content statistics for XML documents per category

Hybrid XML Repository

- No existing general technique effective for any input data
  - using general method only if necessary
- Identification of data patterns
  - frequent parts to be processed specifically
  - preserving updatability
  - XML Schema exploitation
- Numbering schema integration



XML Fragments

- Features of Patterns
  - Frequent usage in real XML documents
  - Apparent meaning/purpose
  - Existence of effective processing method
  - Apparent typical updates and their possible
  - Effective processing
  - Easy recognition
- Fragment categorization
  - known and static (path summary schema)
  - known and finite (path summary schema)
  - mapped to relations (bubble node)
  - mapped to XML-aware text (buble node)
  - unknown or possibly infinite (ORDPATHs like schema)

Adaptability

- Continuous changes should not affect efficiency adversely
- Invocation
  - fragment insertion
  - document insertion
  - query processing
  - automatically maintained background process
- Open issues:
  - similarity function
  - query adaptation
  - transactions

Real XML Documents - Conclusions

- Amount of tagging dominates size of document
- XML Documents are shallow
  - 95% of documents has < 13 max depth,
  - average is about 5
- Highest amounts of elements, attributes, text nodes and mixed contents are at first levels
  - rapid decrease in higher levels (depths)
- Data are regular
  - data-centric documents can often even described by (fairly simple) relational or shallow relational patterns
  - document-centric XML data also contain significant number of patterns
- Most documents use some kind of standard schema

Real XML Documents - Conclusions

- Recursion
  - occurs quite often (doc ~ 43%, ex ~ 64%)
  - the number of recursive elements is low, though
  - it is simple, depth, branching and ed-pair distance is always less than 10
  - the most common type of recursion is linear and pure recursion
  - schemes specify the most general type of recursion
- Mixed contents
  - relatively high usage in document/exchange
  - low usage in data-centric documents
  - mostly simple mixed contents
  - depth is on average less than 10

Thank you

See full text version for references.