

Compressed serialization of semi-structured data

Espen Højsgaard

Rune Højsgaard

28. februar 2005

Introduktion

XMLStore [3] er den centrale komponent i PLAN-X projektet. XMLStore er et applikations-konfigurerbart, distribueret, mobilt persistenslag til opbevaring af semi-strukturerede data (XML dokumenter i særdeleshed) baseret på en værdiorienteret dokumentmodel og grænseflade. XMLStore detekterer og eliminerer automatisk isomorfe undertræer og repræsenterer træer som maksimalt sammensmeltede, orienterede, kredsløse grafer (DAGs¹). I den nuværende implementation bruger XMLStore placeringsuafhængige referencer af fast størrelse (MD5 hashkoder) for kanter og element-tags og tekstdata bliver gemt naivt som ukomprimerede strenge. Endvidere udnytter XMLStore endnu ikke effektiv repræsentation af grammatisk specificerede dokumentunderklasser (som f.eks. ved DTD- eller XMLSchema-specifikationer).

Komprimering af lineær (strengbaseret) data kan konceptuelt opfattes som en firetrin-sproces:

1. Parse inddata til et træ med korte strenge ved bladende og sammensætningsoperatoren ved knuderne. Ved serialiserede repræsentationer af træstruktureret data kan normal parsing bruges.
2. Reducer træet til en DAG, ved at identificere alle strukturelt isomorfe knuder i træet. Dette svarer til at detektere delte undertræer, hvilket klares automatisk af XMLStore.
3. Udfør Huffmankodning på DAGen så knuder med flest indkanter kodes med de korteste binære strenge.
4. Serialisér (udskriv som streng).

Projektet har til hensigt at designe og implementere 3. og 4.

¹directed acyclic graphs

Problemformulering

Projektets formål opsummeres i følgende problemformulering:

Er det muligt at udnytte XMLStore til at komprimere XML-dokumenter lige så hurtigt og pladseffektivt som XMill [5]?

Uddybning og afgrænsning

De fire trin i foregående afsnit udgør tilsammen en XML-kompressor. Det centrale mål i dette projekt er at analysere den resulterende XML-kompressor og at sammenligne den med andre kompressorer mht. brugbarhed, præstation (tids- og pladsforbrug under komprimering og dekomprimering) og kompressionsratio. Der skal sammenlignes med både ordinære strengbaserede kompressorer og XMill, en XML-kompressor der opnår væsentlig bedre kompressionsratioer for dataorienterede XML-dokumenter end gzip og tilsvarende programmer.

Projektets hovedmål er:

1. at designe, implementere og evaluere en pladseffektiv serialisering af data gemt i et XMLStore og tilsvarende deserialisering til XMLStore, baseret på ovenstående idéer. Dette svarer grundlæggende til:
 - at anvende tekstkomprimering på tekstdata og tag-navne.
 - at kode knuderne i grafen binært baseret på deres indvalens.
2. Sammenlign kompressionsratioer og komprimerings-/dekomprimeringshastigheder for XML-dokumenter eksperimentelt og analytisk med både XMill og XOP [2], så vidt muligt.

I det omfang tiden tillader det, vil det endvidere være interessant at undersøge følgende i nævnte rækkefølge:

3. Udvid punkt 1. med idéerne fra XMill. Navnlig skal indholdet af elementer med samme tag komprimeres sammen og separat fra indholdet af andre elementer med andre tags. Dette er for at udnytte sammenhængen mellem tag-navn og elementets indhold. Dette *kan* ødelægge navigerbarheden af de komprimerede data.
4. Tilpas komprimeringsalgoritmen, så det er muligt at navigere komprimeret XMLStore-data uden forudgående deserialisering.
5. Tilpas komprimeringsalgoritmen til at udnytte en fælles grammatisk struktur for en klasse af XMLStore-data (som f.eks. givet i form af en DTD eller et XMLSchema).

Arbejdsopgaver

- Bliv bekendt med XMLStore og værdiorienteret programmering generelt ved at læse [4].
- Find, læs og analyser litteratur om teknikker til at komprimere træstruktureret data; Specifikt studeres XMill nøje forudne XOP.
- Bliv bekendt med XMLStore-implementationen [1].
- Design et kompressionsformat for XMLStore-data (Document Value Model, DVM) baseret på idéerne i delmål 1.
- Implementér og test et program oven på XMLStore til serialisering/deserialisering af DVM, der bruger ovennævnte format.
- Evaluer og sammenlign kompressionsratioer og ydelse (tid og pladsforbrug) for ovennævnte program med XMill, XOP og gzip (og evt. andre kompressore).
- Overvej at integrere førsteordens Markov-model-idéerne (sandsynlighed for det næste symbol er afhængig af forrige symbol) fra XMill. Hvordan forbedrer dette kompressionen? Hvordan påvirkes navigerbarheden? (XMill understøtter ikke navigerbarhed, da det kræver en forudgående deserialisering). Er det muligt at bevare navigerbarheden og samtidig udnytte at fordelingen af et elements indhold afhænger af elementets tag?
- Udarbejd en arbejdsplan for delmål 4 og 5 og for rapportskrivning.
- Deltag i PLAN-X møder.

Tidsplan

28. februar Synopsis afleveres.

3. marts Forsvar af synopsis.

1. juni Undersøgelser, design og implementation er færdige. Herfra arbejdes udelukkende på rapport.

21. juni Rapport afleveres.

24. juni Rapport fremlægges.

Litteratur

- [1] Plan-X project page. <http://www.plan-x.org/>.
- [2] XML-binary Optimized Packaging. <http://www.w3.org/TR/xop10/>.
- [3] XMLStore. <http://www.plan-x.org/xmlstore/>.
- [4] Kasper Bøgebjerg Pedersen and Jesper Tejlgaard Pedersen. Value-oriented XML Store. Master's thesis, ITU and DTU, 2002.
- [5] Suciu and Liefke. XMill: An efficient compressor for XML Data, 1999.