**SIZE--**

**Thomas Dooms, Ward Gauderis, Mano Marichal**

UAntwerpen

België

[thomas.dooms@student.uantwerpen.be](mailto:thomas.dooms@student.uantwerpen.be) [ward.gauderis@student.uantwerpen.be](mailto:ward.gauderis@student.uantwerpen.be) [mano.marichal@student.uantwerpen.be](mailto:mano.marichal@student.uantwerpen.be)

Wanneer de vraag naar toepassingen van context-vrije grammatica’s wordt gesteld, gaan de gedachten bijna onmiddellijk naar parsers en compilers, een vrij nauw applicatiegebied. De bedoeling van het SIZE-- project is om een nuttige applicatie te ontwikkelen die gebruik maakt van de automatentheorie en die het nut van CFG’s in andere domeinen aantoont.

Meestal worden grammatica’s toegepast om het concept van een taal (denk aan programmeertalen en opmaaktalen) op een logische manier te beschrijven. Maar er bestaan ook algoritmes om juist het omgekeerde te doen: om uit een reeds bestaande taal een grammatica af te leiden. Het zijn net die algoritmes waarvan grammar-based compressie gebruik kan maken.

Door deze gedachtegang te volgen is dit projectvoorstel tot stand gekomen. SIZE-- is een tool om bestanden lossless te comprimeren door gebruik te maken van CFG’s, deze op te slaan in een door ons ontwikkeld binair compressieformaat en terug te kunnen extraheren.

Om een bestand te comprimeren moet een taal, namelijk de taal met één string, het inputbestand, omgezet worden in een zo klein mogelijke CFG. Een minimale CFG is een CFG met zo weinig mogelijk symbolen aan de rechterkant van de producties en liefst met ook zo weinig mogelijk producties. Het is aangetoond dat de creatie van een dergelijke CFG voor een taal een NP-hard probleem is. Maar er bestaan verschillende algoritmes die proberen de minimale CFG te benaderen.

Elk compressiealgoritme heeft zijn voor- en nadelen en is daardoor meer geschikt in bepaalde situaties. Dit komt doordat er bij compressie verschillende belangrijke factoren meespelen zoals de tijd- en ruimtecomplexiteit van de compressie (en extractie) en de compressieratio.

Het valt op te merken dat de meeste algoritmes die in dit project van pas komen gebruik maken van straight-line grammatica’s, grammatica’s waarvan de taal bestaat uit exact één string. Deze grammatica’s zullen vaak nog worden opgeslagen in de Chomsky-normaalvorm. De grammar-based compressiealgoritmes die gebruikt zullen worden zijn: het Re-pair, Sequitur, LZW, LCA (online en offline), Bisection en Sequential algoritme. Gevolgd door eventueel het (ingewikkelde) GLZA algoritme indien we tijd over hebben.

De bekomen CFG wordt zo klein mogelijk opgeslagen in het door ons ontworpen binaire bestandsformaat “pal”, staande voor “Public Archive for Losslessness”. Hierbij zal het Huffman algoritme voor variable-width encoding gebruikt worden om de variables en terminals van de CFG op te slaan met een minimale grootte.

Het voordeel van de meeste grammar-based compressiealgoritmes is dat ze een zeer hoge compressieratio hebben en erg snel kunnen worden geëxtraheerd. Bij de extractie moet immers enkel de yield van de unieke parse tree van de CFG berekend worden.

Het doel van dit project is niet alleen aantonen dat compressie met CFG’s mogelijk is, maar ook echt van nut kan zijn. Door gebruik te maken van multithreading zal getracht worden de compressie te versnellen en competitiever te maken. Ook zullen de verschillende algoritmen aan de hand van visualisatie en metingen met elkaar vergeleken worden. Met deze functionaliteit hopen we de gebruiker inzicht te geven in wat er achter de schermen gebeurt.

Het project zal in 3 grote fases verlopen: in de eerste fase wordt het vaste pal-formaat en de nodige infrastructuur errond ontwikkeld, zodat in fase twee de verschillende grammar-based compressiealgoritmes orthogonaal kunnen worden geïmplementeerd zonder code-duplicatie. Tenslotte zal er rond die API een interface voor de gebruikers worden ontworpen en zullen de algoritmes en hun verschillen worden gevisualiseerd.

Omdat er allerlei factoren meespelen in het compressieproces, hebben we veel vrijheid in de compressiemethodes. Dat is de rede dat er in fase 2 vrij veel algoritmes geïmplementeerd zullen worden. Dit vereist veel lees- en opzoekwerk waardoor het project gegarandeerd de volledige 75 werkuren per persoon zal vergen.

SIZE-- is volgens het team een erg verbredend project dat draait rond een onderwerp dat elk van de groepsleden erg interesseert. Het biedt ons de kans om CFG’s op een unieke manier te benutten. We willen met ons project meer inzicht geven in de verschillende grammar-based compressiealgoritmes en hun kwaliteiten.

Daarom is SIZE-- naar onze mening platina waard.