Ingeniare. Revista Chilena de Ingeniería
ISSN: 0718-3291
facing@uta.cl
Universidad de Tarapacá
Chile

Amaya Robayo, Fredy Ángel Miguel; Murillo Fernández, Edwin Andrés
ESTUDIO ESTADÍSTICO DEL NÚMERO DE REGLAS RESULTANTES AL TRANSFORMAR UNA GRAMÁTICA LIBRE DE CONTEXTO A LA FORMA NORMAL DE CHOMSKY
Ingeniare. Revista Chilena de Ingeniería, vol. 18, núm. 2, agosto, 2010, pp. 183-186
Universidad de Tarapacá
Arica, Chile

ESTUDIO ESTADÍSTICO DEL NÚMERO DE REGLAS RESULTANTES AL TRANSFORMAR UNA GRAMÁTICA LIBRE DE CONTEXTO A LA FORMA NORMAL DE CHOMSKY

STATISTICAL STUDY OF THE NUMBER OF RESULTING RULES WHEN TRANSFORMING A CONTEXT-FREE GRAMMAR TO CHOMSKY NORMAL FORM

Fredy Ángel Miguel Amaya Robayo[1] **Edwin Andrés Murillo Fernández**[1]

Recibido 22 de septiembre de 2009, aceptado 13 de julio de 2010
Received: September 22, 2009 Accepted: July 13, 2010

RESUMEN

Es un hecho conocido que toda gramática libre de contexto puede ser transformada a la forma normal de Chomsky de tal forma que los lenguajes generados por las dos gramáticas son equivalentes. Una gramática en forma normal de Chomsky (FNC) tiene algunas ventajas, por ejemplo sus árboles de derivación son binarios, la forma de sus reglas más simples etc. Por eso es siempre deseable poder trabajar con una gramática en FNC en las aplicaciones que lo requieran. Existe un algoritmo que permite transformar una gramática libre de contexto a una en FNC; sin embargo, la cantidad de reglas generadas al hacer la transformación depende del número de reglas en la gramática inicial así como de otras características. En este trabajo se analiza desde el punto de vista experimental y estadístico, la relación existente entre el número de reglas iniciales y el número de reglas que resultan luego de transformar una gramática libre de contexto a la FNC. Esto permite planificar la cantidad de recursos computacionales necesarios en caso de tratar con gramáticas de alguna complejidad.

Palabras clave: Reconocimiento de formas, teoría de autómatas, modelos de lenguaje, gramáticas formales, lenguajes formales.

ABSTRACT

It is well known that any context-free grammar can be transformed to the Chomsky normal form so that the languages generated by each one are equivalent. A grammar in Chomsky Normal Form (CNF), has some advantages: their derivation trees are binary, simplest rules and so on. So it is always desirable to work with a grammar in CNF in applications that require them. There is an algorithm that can transform a context-free grammar to one CNF grammar, however the number of rules generated after the transformation depends on the initial grammar and other circumstances. In this work we analyze from the experimental and statistical point of view the relationship between the number of initial rules and the number of resulting rules after transforming. This allows you to plan the amount of computational resources needed in case of dealing with grammars of some complexity.

Keywords: Pattern recognition, automata theory, language modelling, formal grammars, formal languages.

INTRODUCCIÓN

Una gramática libre de contexto (GLC) es un mecanismo para generar eficientemente lenguajes formales [6], los cuales son sistemas matemáticos usados como modelos teóricos de computación que sirven de apoyo a algunos sistemas de reconocimiento de patrones tales como: reconocimiento de la voz, traducción automática [6] y otros. Las gramáticas también han sido de gran utilidad en el diseño de compiladores, lenguajes de programación para computadoras y como modelo sintáctico de lenguajes naturales [6, 7].

Definición 1. Una gramática formal es una 4-tupla $G = (N, \Sigma, P, S)$ donde:

- Σ es un alfabeto (denominado conjunto de símbolos terminales).
- $X \to \alpha$ es un conjunto finito denominado conjunto de símbolos no terminales, $\Sigma \cap N = \varnothing$.
- P es un conjunto finito de reglas, llamado conjunto de reglas de producción.
- S es el símbolo inicial de la gramática, $S \in N$.

[1] Departamento de Matemáticas. Universidad del Cauca. Popayán, Colombia. E-mail: famaya@unicauca.edu.co; emurillo@unicauca.edu.co

Forma de las reglas. El conjunto finito de producciones o reglas de producción, que representan la definición recursiva de un lenguaje, está formado por expresiones que se escriben en la forma $\alpha \to \beta$ y se leen: "α deriva en β", donde: el lado izquierdo de una producción (antes del símbolo $\to$) se denomina cabeza de la producción o antecedente, en este trabajo se usará antecedente y es un elemento de $(\Sigma \cup N)^+$. Las expresiones $(\Sigma \cup N)^+$, $(\Sigma \cup N)^*$ definen el conjunto de todas las cadenas de longitud mayor o igual que 0 y mayor o igual que 1 (respectivamente) que se pueden formar con elementos de $\sum \cup N$.

- El símbolo de producción: $\to$.

- El lado derecho de una producción es una cadena de $(\Sigma \cup N)^*$. Esta cadena, llamada cuerpo o consecuente de la producción, representa una forma posible de construir cadenas en el lenguaje. Para hacerlo, si el antecedente aparece en una cadena de $(\Sigma \cup N)^*$ éste puede ser reemplazado por el consecuente.

En adelante utilizaremos las siguientes convenciones con respecto a las gramáticas:

1. Si $\alpha \to \beta$ es una regla de producción que se aplica a la cadena $\gamma\alpha\delta$ para derivar la cadena $\gamma\beta\delta$, esto se denota $\gamma\alpha\delta \Rightarrow \gamma\beta\delta$.

2. Supóngase que $\alpha_1, \alpha_2, ..., \alpha_m$, son cadenas de $(\Sigma \cup N)^*$, $m \geq 1$ y $\alpha_1 \Rightarrow \alpha_2 \Rightarrow \alpha_3 \Rightarrow, ..., \Rightarrow \alpha_{m-1} \Rightarrow \alpha_m$. Entonces se dice que $\alpha_1 \overset{*}{\underset{G}{\Rightarrow}} \alpha_m$ o α_1 deriva en α_m en la gramática *G*.

Las gramáticas se clasifican mediante la jerarquía de Chomsky en cuatro tipos [3]; dentro de estos es de particular importancia el estudio de las GLC. Al utilizar una GLC para generar un lenguaje existen dos problemas básicos: el primero, consiste en que si se permite en el lenguaje el uso de cadenas de longitud nula, esto puede hacer que el proceso para generar una determinada cadena nunca termine. El segundo, la existencia de cierto tipo de reglas en la gramática; unas de ellas denominadas unitarias y las otras inútiles, las cuales hacen que el trabajo para producir una cadena pueda ser excesivo, pues las unitarias no contribuyen a la generación del lenguaje ya que solo hacen el reemplazo de una variable por otra mas no por un símbolo terminal que es lo que se necesita para producir cadenas; por otro lado, las reglas inútiles no contribuyen en el proceso de generar alguna cadena pues dichas reglas nunca se emplean en la obtención de las cadenas.

Afortunadamente existe un procedimiento que permite transformar la GLC a formas más simples que no presenten las dificultades mencionadas y tal que el lenguaje generado por la nueva gramática sea equivalente al lenguaje generado por la gramática inicial. La forma más utilizada es la forma normal de Chomsky (FNC), en la cual la forma de las reglas es muy simple y manejable desde el punto de vista computacional.

Definición 2. Una gramática libre de contexto (GLC) o de tipo 2. Se caracteriza porque las reglas son de la forma $X \Rightarrow \alpha$, donde *X* es un no terminal o variable y α es una cadena que puede contener no terminales y símbolos terminales. Los lenguajes que estas gramáticas producen se llaman lenguajes libres de contexto (LLC).

Definición 3. El lenguaje generado por *G* es el conjunto *L(G)*, definido como $L(G) = \{x \in \Sigma^* / S \overset{*}{\Rightarrow} x\}$.

Definición 4. (Forma normal de Chomsky). Sea $G = (N, \Sigma, P, S)$ una gramática. Se dice que *G* está en forma normal de Chomsky si el conjunto de reglas de producción *P* está constituido por reglas de la forma $A \to BC$ o $A \to a$ donde $A, B, C \in N$ y $a \in \Sigma$.

Para abordar el tema de este trabajo cabe preguntar: si la gramática inicial tiene $|P|$ reglas, al transformarla a la FNC, ¿cuántas reglas tendrá la gramática final? La pregunta tiene importancia pues al estudiar el teorema que transforma una GLC a la FNC [6] se concluye que no es posible determinar la cantidad de reglas resultantes en FNC a partir de la GLC inicial. Esto motiva la realización de un estudio estadístico que permita estimar un modelo que establezca la relación entre el número inicial de reglas y el final. Tal modelo permite estimar de antemano la cantidad de recursos necesarios para almacenar la gramática resultante en las aplicaciones computacionales.

ANÁLISIS ESTADÍSTICO EXPERIMENTAL

En el desarrollo de esta sección se realizará un análisis de regresión, el cual permitirá plantear un modelo matemático que establece la relación entre el número de

reglas de la GLC y el número de reglas de la gramática en FNC.

Metodología experimental

Inicialmente se tiene un conjunto de terminales Σ y un conjunto de no terminales *N*. A partir de éstos se generan $|P|$ reglas de manera aleatoria, reglas de la forma $X \to \alpha;\ X \in N,\ \alpha \in (N \cup \Sigma)^*$.

Así se obtiene una gramática $G = (N, \Sigma, P, S)$ con $|P|$ reglas. Esta gramática se transforma a la FNC, obteniéndose una nueva gramática G' en FNC, $G' = (N, \Sigma', P', S)$ con P' reglas y tal que $L(G) = L(G')$.

Sea $|P| = n$, el proceso descrito anteriormente se repite 5.000 veces, es decir, se generan 5.000 GLC y se transforman a FNC; se anotan los números de reglas resultantes en cada una de las 5.000 transformaciones. Por último, se toma de los 5.000 datos como valor representativo el número promedio de reglas y el peor de los casos (el que más reglas finales en FNC produjo). Este experimento se repite para $|P| = n = 100$ hasta n = 3.000 variando de 100 en 100.

En la siguiente sección se muestran los resultados.

Análisis estadístico

Con los datos obtenidos de los experimentos se realizó un análisis de regresión y correlación para los dos casos mencionados anteriormente es decir:

1. peor caso.
2. promedio.

En los dos casos las variables independiente y dependiente para el análisis de regresión se definen respectivamente como $x = |P|$ y $y = |P'|$.

En la Figura 1 en color azul se presentan los datos originales del peor caso y en color rojo los del caso promedio.

Se observa en la Figura 1 que a partir de 300 reglas iniciales el comportamiento para el caso promedio tiene tendencia lineal. Mientras que el peor caso muestra un comportamiento cuadrático.

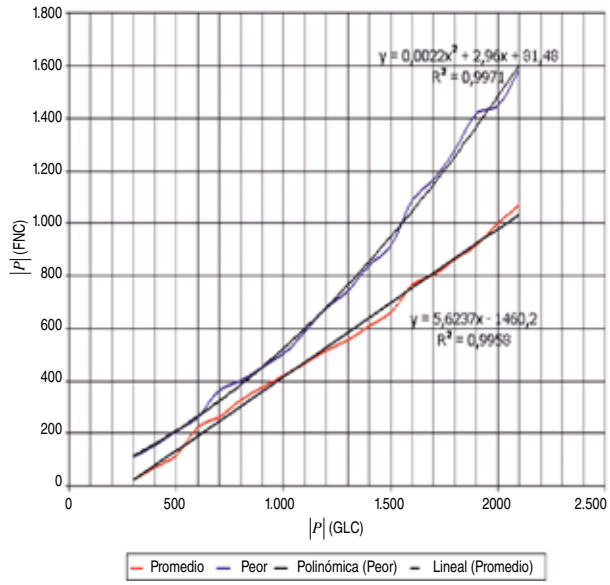

Figura 1. Relación entre el número de las reglas de una GLC y el número de reglas al ser transformadas a la forma normal de Chomsky.

Esta apreciación ha motivado hacer un estudio de regresión y correlación de los dos casos.

Para el caso promedio en la misma Figura 1 se ha agregado una recta de regresión lineal, la cual aparece de color negro, y al lado de ella se muestra el valor del coeficiente de determinación que indica que la relación lineal es buena entre el número inicial de reglas de la GLC y el número final de reglas en forma normal de Chomsky. En nuestro caso la relación es del 99,58%, lo cual significa que la relación es estrecha y creciente. La recta de regresión lineal tiene como modelo

$$y = 5,6237x - 1460,2 \qquad (1)$$

Donde *y* en la ecuación (1) representa el número aproximado de reglas en forma normal de Chomsky y *x* es el número de reglas de la gramática inicial. Así, el modelo nos permite estimar la cantidad de memoria que se requiere cuando se va a transformar una gramática libre de contexto a la FNC.

Para el primer caso, en la parte superior de la Figura 1, se muestra cómo evoluciona en el peor de los casos el número de reglas resultantes en forma normal de Chomsky en función del número de reglas iniciales de la gramática libre de contexto. Se ha aproximado la curva mediante un modelo de regresión cuadrática, por:

$$y = 0,0022x^2 + 2,96x + 81,48 \qquad (2)$$

Donde *y* en la ecuación (2) es aproximadamente el número de reglas en forma normal de Chomsky y *x* es el número de reglas de la gramática libre de contexto inicial. Para este caso la relación es del 99,71%, lo cual muestra una buena relación entre las variables y con tendencia creciente.

CONCLUSIONES

En este trabajo se realizó un análisis estadístico en el que se ha establecido la existencia de una relación lineal entre el número inicial de reglas de la GLC y el número final de reglas en FNC en el caso promedio definido en el texto. También se halló, en el peor caso, una relación cuadrática entre el número inicial de reglas de una GLC y el número final de reglas en FNC. Los modelos matemáticos se muestran en (1) y (2) y representan tanto la relación lineal como la cuadrática; se observa en la Figura 1 el ajuste del modelo a los datos con un alto grado de confianza. Esto significa que dada una gramática GLC arbitraria de antemano podemos estimar con un alto grado de probabilidad el número final de reglas que van a resultar luego de transformarla a una gramática equivalente en FNC. Así, se dispone a partir de ellos de un método para estimar los recursos computacionales de espacio en memoria cuando se va a transformar una GLC a la FNC.

REFERENCIAS

[1] F.A. Amaya Robayo. "Algunos aportes a los modelos de lenguaje de máxima entropía de frase completa". Tesis para optar el grado de doctor. Universidad Politécnica de Valencia. Valencia, España. 2001.

[2] A.F. Cárdenas. "Ciencias de la computación". Limusa-Wiley S.A. 1972.

[3] N. Chomsky. "Aspects of the Theory of Syntax". The MIT Press. Cambridge, Massachusetts, EEUU. 1965.

[4] H. Ney. "Stochastic Grammars and pattern recognition". In P. Laface and R. de Mori (Eds.). Speech Recognition and Understanding. Springer Verlag, pp. 319-344. 1992.

[5] H. Ney. "Dynamic programing parsing for context-free grammars in continuous speech recognition". IEEE Transactions on Signal Processing. Vol. 39, Issue 2, pp. 336-340. February, 1991.

[6] J.E. Hopcroft, R. Motwani y J.D. Ullman. "Introducción a la teoría de autómatas, lenguajes y computación". Addison-Wesley, 2nd Edition. 2002.

[7] L. Miclet. "Structural Methods in Pattern Recognition". North Oxford Academic Press. 1986.

[8] E. Murillo. "Gramáticas libres de contexto, aproximación computacional". Tesis para optar el grado de matemático. Universidad del Cauca. Popayán, Colombia. 2008.

[9] E. Murillo. "Reducción de algoritmos GLC". XIII Congreso de la escuela regional de matemáticas. Pereira, Colombia. Septiembre 2006.

[10] A. McFarlane Mood, F.A. Graybill and D.C. Boes. "Introduction to the Theory of Statistics". McGraw-Hill, 3rd Edition. 1974.

[11] J.A. Sánchez. "Estimación de gramáticas incontextuales probabilísticas y su aplicación en modelización del lenguaje". Tesis para optar el grado de doctor. Universidad Politécnica de Valencia. Valencia, España. 1999.