Procesamiento del Lenguaje Natural, num. 30 (2003), pp.121-122 recibido 1-03-2003; aceptado 20-05-2003

Análisis Eficaz de Gramáticas de Cláusulas Definidas

David Cabrero Souto
Departamento de Computación, Universidade da Coruña
Campus de Elviña s/n, 15071 A Coruña
cabrero@udc.es

Resumen: Tesis doctoral en Informática realizada por David Cabrero Souto bajo la dirección del doctor Manuel Vilares Ferro (Universidade da Coruña). El acto de defensa de la tesis tuvo lugar el 19 de octubre de 2002 ante el tribunal formado por los doctores Enric Trillas Ruiz (Universidad Politécnica de Madrid), Miguel Delgado Calvo-Flores (Universidad de Granada), Eric Villemonte de la Clergerie (INRIA, Francia), Antonio Blanco Ferro (Universidade da Coruña) y Jose Luis Freire Nistal (Universidade da Coruña). La calificación obtenida fue Sobresaliente Cum Laude por unanimidad. Se puede obtener más información de la tesis en http://www.grupocole.org/~cabrero/ o contactando directamente con el autor a través de la dirección de correo electrónico cabrero@udc.es.
Palabras clave: Análisis sintáctico, autómatas, gramáticas de cláusulas definidas, análisis parcial, programación dinámica.

Abstract: PhD Thesis in Computer Science written by David Cabrero Souto under the supervision of Dr. Manuel Vilares (Universidade da Coruña, Spain). The author was examined in October 19, 2002 by the commitee formed by Dr. Enric Trillas (Universidad Politécnica de Madrid), Dr. Miguel Delgado (Universidad de Granada), Dr. Eric de la Clergerie (INRIA, Francia), Dr. Antonio Blanco (Universidade da Coruña) y Dr. Jose Luis Freire (Universidade da Coruña). The grade obtained was *Sobresaliente Cum Laude*. Further information is available at http://www.grupocole.org/~cabrero/ or at email address cabrero@udc.es.
Keywords: Automata, dynamic programming, parsing, definite clause grammars, partial parsing.

1. Introducción

El análisis sintáctico ha sido y continúa siendo investigado activamente. La base tecnológica de esta investigación incluye la mejora en las técnicas de análisis sintáctico, conocimiento semántico, motores morfológicos, ...y el desarrollo de *shallow parsers*, término que podríamos hacer corresponder en castellano con *analizador sintáctico superficial*. Destacar también, que dentro del análisis sintáctico, la utilización de formalismos gramaticales valuados es, hoy por hoy, una referencia indiscutible en lo que se refiere a los entornos de procesamiento del lenguaje natural, y en programación lógica, este último representante del paradigma declarativo. La presente tesis doctoral aborda el estudio y desarrollo de técnicas de análisis sintáctico dirigidas, precisamente, al tratamiento de sistemas basados en el análisis de formalismos gramaticales valuados.

En esta tesis se trata el desarrollo de analizadores sintácticos, tanto en lo que respecta a su ejecución eficiente, como a la extensión del dominio de los formalismos gramaticales a los que resultan aplicables. Para ello se han adoptado las siguientes decisiones de diseño:

- Uso de máquinas abstractas o *autómatas* en lugar de la propia gramática. En efecto, un autómata no es más que un dispositivo matemático que permite un tratamiento más eficiente del proceso de análisis.
- Uso de programación dinámica. La presencia de ambigüedades conlleva la repetición de cálculos, y la recursión, problemas de completud operacional. Las técnicas de programación dinámica abordan estos inconvenientes mediante la compartición de cálculos.
- Análisis estático. Con objeto de reducir el espacio de búsqueda, se realiza un análisis previo de la gramática. Los resultados de este análisis permiten mejorar la complejidad tanto temporal como

ISSN 1135-5948

espacial de los algoritmos.

- Indexación del proceso de análisis. La sincronización del proceso mediante la indexación reduce el espacio de búsqueda.
- Compartición de estructuras. La compartición no sólo mejora la complejidad espacial, sino que además permite la implementación eficiente de operaciones básicas como la *unificación*.
- En el caso del análisis del lenguaje natural, uso de un *lexicón* separado. De esta forma se permite el empleo de técnicas mejor adaptadas a su casuística particular, que difiere de la presente en el análisis sintáctico.

2. Estructura de la tesis

La estructura lógica consta de varias partes en las que se aborda de manera incremental los conceptos y técnicas expuestos, de forma que cada parte incluye los resultados de las anteriores. A continuación esbozamos las líneas generales de cada una:

En una primera parte encontramos un recorrido por las técnicas clásicas y otras más actuales relacionadas con el análisis sintáctico, centrándose en las más representativas. El objeto es ofrecer una visión general del estado del arte en el campo que nos ocupa.

En una segunda parte se trata el problema del análisis de *gramáticas independientes del contexto* (GICs). El objetivo es introducir y desarrollar técnicas como los *sistemas de deducción gramatical* (SDGs), *autómatas de pila* (APs) y sus *esquemas de compilación*, y *tabulación*. Estas técnicas se emplearán en el resto de la memoria. En un primer capítulo se introduce el problema del análisis sintáctico de GICs. Se ha optado por estas gramáticas debido a su amplia difusión y conocimiento, de esta forma se establecen los conceptos y terminología empleados en el resto de la memoria. A continuación se presentan los sistemas de deducción gramatical como marco descriptivo común bajo el cual se formulan diversos algoritmos de análisis, mostrando la relación entre ellos, obteniendo unos a partir de los anteriores. Para completar esta parte, se desarrolla la aplicación de los *autómatas de pila* a la construcción de analizadores sintácticos. Asimismo, se introduce la aplicación de las técnicas de tabulación a los autómatas obtenidos para la mejora de su eficiencia. Finalmente se describe el sistema *ice*, de generación de analizadores sintácticos para GICs sin restricciones.

La tercera parte trata el problema del análisis de GCDs. El poder descriptivo de las GICs se muestra insuficiente para algunas aplicaciones como pueda ser el *procesamiento del lenguaje natural*. Como alternativa se presentan formalismos más complejos, entre ellos, las *gramáticas lógicas*. En concreto, optaremos por las GCDs como formalismo gramatical de referencia. Este formalismo goza de un mayor poder descriptivo, pero a cambio se incrementa la complejidad de los analizadores asociados. Se trata la descripción de las GCDs como una generalización de las GICs, la adaptación de los SDGs presentados para dichas gramáticas, la generalización de los APs para la implementación de analizadores, y los mecanismos de tabulación asociados. Finalmente se trata la evolución del sistema *ice* a las GCDs. Seguidamente, se estudian las estructuras de datos empleadas para una implementación eficiente de la operación de unificación, que constituye la base del formalismo de GCDs. A continuación se aborda el tratamiento de *estructuras cíclicas*, con la consiguiente ampliación del dominio de terminación de los analizadores propuestos.

La cuarta parte incluye aquellos aspectos relacionados con la aplicación de las técnicas de análisis sintáctico al problema del procesamiento del lenguaje natural, en la que resultan especialmente importantes los resultados de capítulos anteriores dirigidos a la mejora de la eficiencia, tratamiento de formalismos gramaticales complejos, o la ampliación del dominio de terminación. En relación a la adecuación de los analizadores presentados anteriormente al procesamiento del lenguajes natural, se tratan tres aspectos importantes: la integración con otras herramientas habituales como los *etiquetadores*, el tratamiento de palabras desconocidas y la obtención de análisis sintácticos parciales.

En una última parte, encontramos un capítulo dedicado a la presentación de los resultados obtenidos tras la realización de diversos experimentos sobre las técnicas y algoritmos de análisis sintáctico tratadas en los capítulos anteriores. Por último, se exponen las conclusiones sobre el trabajo realizado y se trazan las posibles líneas a seguir para su mejora y ampliación.