Procesamiento del Lenguaje Natural, num. 30 (2003), pp.119-120 recibido 1-03-2003; aceptado 20-05-2003

Reconocimiento de patrones en bosques compartidos

Francisco José Ribadas Pena
Departamento de Computación, Universidade da Coruña
Campus de Elviña s/n, 15071 La Coruña
ribadas@mail2.udc..es

Resumen: Tesis doctoral en Informática realizada por Francisco J. Ribadas Pena bajo la dirección del doctor Manuel Vilares Ferro (Universidade de Vigo). El acto de defensa de la tesis tuvo lugar el 29 de noviembre de 2002 ante el tribunal formado por los doctores José Luis Freire Nistal (Universidade da Coruña), Alejandro Sobrino Cerdeiriña (Universidade de Santiago de Compostela), Antonio Bahamonde Rionda (Universidad de Oviedo), Richard F.E. Sutcliffe (University of Limerick, Irlanda) y Jean-Cêdric Chappelier (Swiss Federal Institute of Technology, Suiza). La calificación obtenida fue Sobresaliente Cum Laude por unanimidad. Se puede obtener más información de la tesis en http://coleweb.dc.fi.udc.es.
Palabras clave: Reconocimiento de patrones, análisis sintáctico, recuperación y extracción de información, programación dinámica.

Abstract: PhD Thesis in Computer Science written by Francisco J. Ribadas Pena under the supervision of Dr. Manuel Vilares (Universidade de Vigo, Spain). The author was examined in November 29, 2002 by the commitee formed by Dr. José Luis Freire Nistal (Universidade da Coruña, Spain), Dr. Alejandro Sobrino Cerdeiriña (Universidade de Santiago de Compostela, Spain), Dr. Antonio Bahamonde Rionda (Universidad de Oviedo, Spain), Dr. Richard F.E. Sutcliffe (University of Limerick, Ireland) and Dr. Jean-Cêdric Chappelier (Swiss Federal Institute of Technology, Switzerland). The grade obtained was *Sobresaliente Cum Laude*. Further information is available at http://coleweb.dc.fi.udc.es.
Keywords: Pattern matching, parsing, information retrieval and extraction, dynamic programming.

1 Introducción

El reconocimiento de patrones se puede definir como el proceso de encontrar una subestructura objetivo incluida dentro de otra mayor, mediante la comparación de ésta con una representación de aquella en forma de patrón. De este modo, el reconocimiento de patrones constituye un marco prometedor para la construcción de sistemas de gestión y localización de información en donde la estructura de los elementos almacenados sea relevante. Ofreciendo un modo natural para interrogar de forma descriptiva un conjunto de estructuras complejas.

En el caso que nos ocupa, los elementos sobre los que se aplicarán estas técnicas de reconocimiento de patrones tendrán la forma de árboles o bosques, resultado del análisis sintáctico de sentencias escritas en lenguaje natural. Nuestro trabajo se enmarca dentro del estudio de la aplicación de técnicas de procesamiento del lenguaje natural (PLN) en sistemas de recuperación y extracción de información (RI, EI). En concreto, esta tesis está relacionada con la utilización de estructuras sintácticas para mejorar el rendimiento de determinados aspectos de estos sistemas, como la consulta o la indexación.

2 Reconocimiento de patrones en árboles

Desde el punto de vista formal, el reconocimiento de patrones en árboles es una evolución de su homónimo en cadenas de caracteres. Este es un problema ampliamente estudiado, debido principalmente a sus aplicaciones en el campo de los sistemas de RI. Sin embargo, la comparación de árboles es un problema inherentemente más costoso desde un punto de vista computacional. Esto tiene como consecuencia que este tipo de técnicas vea frenada su utilización como mecanismo básico en determinadas aplicaciones prácticas. En esos casos se suelen emplear estrategias menos poderosas semánticamente, pero más eficientes computacionalmente.

ISSN 1135-5948

Así sucede en los sistemas de RI clásicos, basados en el concepto de palabra clave. Sin embargo, en muchos casos, es el modo en que se combinan dichas palabras junto con sus implicaciones semánticas lo que determina la auténtica relevancia de un documento respecto a una consulta. Esto lleva a plantear la necesidad de representaciones más sofisticadas. Diferentes autores han estudiado la posibilidad de utilizar las estructuras que resultan del análisis sintáctico de los documentos para mejorar el proceso de indexación y/o consulta de estos sistemas. Dentro del marco descrito, nuestra aportación está orientada al desarrollo de técnicas eficientes que extiendan los algoritmos clásicos para permitir el reconocimiento de patrones sobre bosques de árboles altamente ambiguos.

La aproximación clásica al problema del reconocimiento de patrones en árboles se basa en el concepto de distancia de edición. La idea básica es cuantificar la similaridad entre dos estructuras tomando como medida el coste de la transformación de una de ellas en la otra. Para ello se define un conjunto de operaciones de edición, que permiten modificar los elementos de una estructura, y se asocia a cada una de ellas un coste numérico.

De cara a la aplicación práctica de estas técnicas, es también interesante disponer de un mecanismo de reconocimiento de patrones lo más flexible posible. Es conveniente manejar mecanismos de reconocimiento aproximado, que permitan una especificación vaga de los patrones, de modo que los usuarios puedan omitir detalles estructurales irrelevantes o desconocidos. En nuestro caso hemos optado por dar soporte a la inclusión de símbolos VLDC (por *Variable Length Don't Care*) que permiten especificar de forma aproximada los árboles patrón que se desean identificar.

3 Análisis sintáctico y reconocimiento de patrones

En el caso que nos ocupa, nuestro objetivo ha sido trasladar la aproximación clásica del cálculo de distancias de edición al tipo de bosques compartidos resultantes del análisis de una sentencia en lenguaje natural, generalmente ambigua, generado por el sistema ICE (*Incremental Contex-free Environment*). Este es un generador de analizadores sintácticos incrementales para gramáticas de contexto libre sin restricciones, basado en el concepto de interpretación tabular de autómatas con pila. La salida de los analizadores generados por ICE tiene la forma de un grafo Y-O, que representa de forma compacta y sin redundancia todos los posibles árboles de análisis asociados a la sentencia de entrada. Se ha pretendido, además, sacar el mayor partido posible de la compartición de estructuras ofrecida por ICE, para evitar la repetición de cálculos.

Para llevar a cabo esta integración hemos realizado una revisión detallada de los algoritmos clásico de reconocimiento de patrones en árboles, centrándonos tanto en aspectos relativos a su eficiencia computacional, como en sus capacidades en cuanto a potencia expresiva. Para realizar eficazmente su integración en ICE ha sido necesario analizar el tipo de representación sintáctica utilizada, identificando los factores que determinan la compartición de estructuras en el análisis de sentencias ambiguas y cómo afectan al mecanismo de reconocimiento de patrones. Esto nos ha permitido modificar las propuestas originales para adaptarlas a las peculiaridades de las estructuras sintácticas manejadas. Las modificaciones propuestas son compatibles con la posibilidad de añadir capacidades de reconocimiento de patrones más sofisticadas. En concreto, se plantean una serie de extensiones del mecanismo de reconocimiento para incluir características orientadas específicamente al procesamiento de árboles de análisis sintáctico y en concreto, a su aplicación en sistemas de RI, como por ejemplo el empleo de símbolos VLDC o la inclusión de variables de extracción.

Por último, hemos realizado una evaluación de las modificaciones propuestas, tanto desde el punto de vista de su coste computacional como desde sus posibilidades de utilización por parte del usuario. Para dicha evaluación hemos empleado un conjunto de gramáticas altamente no deterministas, que ponen de manifiesto una mejora importante del rendimiento en el caso del reconocimiento de patrones en bosques compartidos.

Pese a estas optimizaciones, de cara a su aplicación directa en sistemas de RI sin restricciones, el coste de nuestra aproximación es claramente superior al de las técnicas clásicas basadas en reconocimiento de cadenas de caracteres. Sin embargo, confiamos en sus posibilidades de aplicación en entornos restringidos y muy especializados, donde se manejen volúmenes de datos no excesivamente altos y se exija una alta exactitud.