



Apr 12th, 9:00 AM - 10:00 AM

P-Fastus: Information Extraction System Implemented in Prolog (SICStus)

Rajen Subba

Illinois Wesleyan University

Hans-Joerg Tiede, Faculty Advisor

Illinois Wesleyan University

Follow this and additional works at: <http://digitalcommons.iwu.edu/jwprc>

Subba, Rajen and Tiede, Faculty Advisor, Hans-Joerg, "P-Fastus: Information Extraction System Implemented in Prolog (SICStus)" (2003). *John Wesley Powell Student Research Conference*. 15.
<http://digitalcommons.iwu.edu/jwprc/2003/posters/15>

This Event is brought to you for free and open access by The Ames Library, the Andrew W. Mellon Center for Curricular and Faculty Development, the Office of the Provost and the Office of the President. It has been accepted for inclusion in Digital Commons @ IWU by the faculty at Illinois Wesleyan University. For more information, please contact digitalcommons@iwu.edu.

©Copyright is owned by the author of this document.

Poster Presentation P41

**P-FASTUS: INFORMATION EXTRACTION SYSTEM
IMPLEMENTED IN PROLOG (SICStus)**

Rajen Subba and Hans-Joerg Tiede*
Department of Computer Science, Illinois Wesleyan University

P-FASTUS is an Information Extraction (IE) system developed in SICStus Prolog based on the implementation of FASTUS. FASTUS is an IE system developed by Stanford Research International uses a cascade of finite state automatons.

A vast majority of the information held by businesses, government agencies and individuals alike are stored in text files. With the advent of the internet, the amount of textual information in the form of natural languages has been growing exponentially. Searching for documents containing relevant information on the web has become a fairly daunting task. Reading through thousands of documents to obtain the information that you require can be cumbersome. In order to address this issue, researchers have been developing Information Extraction systems using the techniques of Natural Language Processing (NLP).

The goal of Information Extraction is to extract from a set of documents, prominent facts about pre-specified types of events, entities or relationships. P-FASTUS is a system that extracts pre-specified information such as the name of the company, location and the position being advertised from "Job Postings" in text files. The system, like FASTUS, is composed of different levels of processing that are developed using Finite State Automatons.

Finite State Automatons are ideal machines not bound by any physical constraints that are composed of one or more states. The machine moves from one state to another based on the input. The movement is governed by a transition function at each state. When the entire input is read and the machine either stops at what is known as an accepting state or state that rejects the input. For the purposes of Information extraction Finite State Automatons are used to approximate finite-state grammars which are then used for pattern matching of specific linguistic constructs that contain the information desired to be extracted.

Most of the IE systems were implemented in Lisp and C and none in Prolog despite the fact that Prolog's features make it a language that is more suitable for NLP. SICStus is a version of Prolog that supports constraint programming capabilities. The goal of this project was to implement FASTUS in a constraint logic programming language and assess possible advantages of implementing an IE system in such a language.