Expert Systems with Applications 40 (2013) 3266–3281

Contents lists available at SciVerse ScienceDirect

Expert Systems with Applications

journal homepage: www.elsevier.com/locate/eswa

Ontology-based information extraction of regulatory networks from scientific articles with case studies for *Escherichia coli*

Antonio Moreno [a], David Isern [a,*], Alejandra C. López Fuentes [b]

[a] Universitat Rovira i Virgili, Department of Computer Science and Mathematics, ITAKA research group, Avda. Països Catalans, 26, 43007 Tarragona, Catalonia, Spain
[b] Programa de Genómica Computacional, Centro de Ciencias Genómicas, Universidad Nacional, Autónoma de México, Cuernavaca, Morelos, México

ARTICLE INFO

Keywords:
Ontology-based information extraction
Knowledge representation
Regulatory networks

ABSTRACT

The amount of scientific papers in the Molecular Biology field has experienced an enormous growth in the last years, prompting the need of developing automatic Information Extraction (IE) systems. This work is a first step towards the ontology-based domain-independent generalization of a system that identifies *Escherichia coli* regulatory networks. First, a domain ontology based on the RegulonDB database was designed and populated. After that, the steps of the existing IE system were generalized to use the knowledge contained in the ontology, so that it could be potentially applied to other domains. The resulting system has been tested both with abstract and full articles that describe regulatory interactions for *E. coli*, obtaining satisfactory results.

1. Introduction

The Web is a vital tool in our Information Society. It has grown exponentially through the 1990s and into the 21st century and it has expanded to serve hundreds of millions of users and a multitude of purposes in all parts of the world. The Web carries a vast range of information resources that provide general information which can be easily found on any subject with the simple click of the mouse. Thousands of new Web pages appear every day: news, magazines, and journal articles are constantly being created online. All these trends result in an enormous amount of text available in digital form; however, these repositories of text are mostly untapped resources of information, and identifying specific desired information in them becomes a difficult task.

The Web is drowning us with an enormous amount of pages of information related to all domains of discourse. Recently, the fields of Molecular Biology and Medicine have enjoyed an explosive development; as a result, the number of biomedical journal articles has grown exponentially, and an urgent need on the part of biologists to find specific information in the ever expanding biological literature has emerged.

Access to full-text articles is difficult because each journal has its own organization, interface and formatting conventions. Fortunately, in the Biology and Medicine fields the abstracts of the scientific papers are collected and indexed in MEDLINE. This repository, hosted at the National Library of Medicine (NLM), is the primary component of PubMed[1] that indexes over 5000 different journals in Medicine and Molecular Biology. It currently contains more than 22 million abstracts and it keeps growing.

Today, it is possible to retrieve articles using queries that can be very complex (including different attributes of the publication, combined with logical operators) and to read their abstracts. With the increasing distribution of journals in electronic format, the full text of any paper will soon be only one click away (Miguel A. Andrade & Bork, 2000). However, the easy access to the literature does not solve the problem of the selection of information. Reading or even browsing all of those publications is something that most researchers will not contemplate. Clearly, there is a need of developing methods for automatic extraction of relevant information from any source of scientific data, especially sources such as scientific literature written in natural language.

Molecular Biology is a branch of Biology which primarily deals with functions, characteristics and structures of mainly three major macro-molecules: DNA, RNA and proteins. Molecular research technologies and developments in Information Technologies have combined to produce a huge amount of information related to the multiple research areas of Molecular Biology such as sequence analysis, genome annotation, analysis of gene expression, analysis of regulation, etc. Several databases have been designed and implemented in order to compile all this information (Benson, Karsch-Mizrachi, Lipman, Ostell, & Wheeler, 2008; Grivell, 2002; Keseler et al., 2005). However, the amount of knowledge in the domain is overwhelming and grows at an unprecedented rate. Much of it

* Corresponding author. Tel.: +34977256563; fax: +34977559710.
E-mail addresses: antonio.moreno@urv.cat (A. Moreno), david.isern@urv.cat (D. Isern), aclopez@ccg.unam.mx (A.C. López Fuentes).

[1] Website: http://www.ncbi.nlm.nih.gov/pubmed (Last access: November 3th, 2012).

0957-4174/$ - see front matter © 2012 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.eswa.2012.12.090

is published and available electronically in the form of scientific texts, therefore, in natural language.

The main hypothesis in this work is that it is necessary to develop improved computer-based tools to aid human experts to extract the facts needed by the research community (Collado-Vides et al., 2009). In the recent years, Information Extraction (IE) has emerged within the Artificial Intelligence community to face this problem. IE systems usually follow a sequence of steps that gradually identify pieces of information, analyze them, and finally represent them in a structured way. This paper presents an evolution of those traditional systems. It shows the current implementation of an IE system to extract regulations on the biological domain using some language-dependent elements, CCG-IE, developed in the UNAM's Centre for Genomic Sciences, and proposes an evolution to add a higher level of expressivity using a widely-used model of knowledge representation (ontologies). The main goal of the work has been to create an ontological structure used to guide the extraction of information. The design and use of this ontology is crucial to create a more powerful system able to extract information from different domains.

The paper explains in detail the design and the main classes of the domain ontology. In this case, due to the complexity of the domain of research, the ontology has been constructed from a large database of gene regulations (RegulonDB). Additionally to this corpus of terms, the ontology requires more elements such as the representation of grammars used to annotate sentences in natural language-based texts. The paper also describes the ontology-based IE system and the tests that have been performed on the current prototype.

The rest of the paper is organized as follows. First of all, Section 2 presents the state of the art on ontology-based IE, paying special attention to the available IE techniques and their application in Molecular Biology. Section 3 describes the domain ontology, which is used later to guide the IE process (Section 4). Section 5 presents some case studies of the current prototype and, finally, Section 6 lists some conclusions and lines of future research.

2. Related work

This section provides an introduction to Information Extraction a classification of the main techniques used in this field (statistical, rule-based and ontology-based) and a general architecture of ontology-based IE systems (Wimalasuriya & Dou, 2010). Then, an outline of five recent biomedical Information Extraction systems is performed, and a comparative table of the analyzed systems with the one proposed in this paper is given.

2.1. Information extraction techniques

Information Extraction is concerned with selecting salient facts about a given topic from a set of documents. Typically, these facts are then entered automatically into a database, which may then be used for further processing. IE is a technology based on natural language analysis. An IE system must be designed to extract the entities and relations appropriate to a specific task (Blaschke, Hirschman, & Valencia, 2002). In general, the core of an IE system is an extractor which processes text; it overlooks irrelevant words and phrases and attempts to home in on entities and the relationships between them (Etzioni, Banko, Soderland, & Weld, 2008). There are several methods to perform IE which can be classified into *statistical* methods and *rule-based* methods, being the later superior to the former in those tasks that require the extraction of relationships (expression of a specific product of a certain gene in a certain tissue, specific function of a protein, etc.), although they are harder to implement (Feldman, Regev, Finkelstein-Landau, Hurvitz, & Kogan, 2002).

Additionally, another approach has recently emerged: ontology-based information extraction (OBIE). Since IE is essentially concerned with the task of retrieving information for a particular domain, specifying the concepts of that domain formally and explicitly through an ontology can be helpful to this process (Wimalasuriya & Dou, 2010).

2.1.1. Statistical information extraction

The statistical approach uses training material which can be annotated (usually through Machine Learning techniques). Several methods are based on the frequency of co-occurrence of words in a large text corpus. The basic elements of text are words, and their frequencies, co-occurrences and lexical features can be used to cluster and classify text, find documents that treat a similar theme or select words that describe a group of documents. For example, M.A. Andrade and Valencia (1998) used the characteristics of word distributions in text clusters to extract significant words. These approaches are limited because words are often ambiguous and refer to more than one object. Moreover, different words can have the same meaning (synonyms) and the same word can be part of constructions with very different meanings.

There are many statistical methods that have been applied for specific tasks in the IE process and have proved to be very effective where there is a large corpora of training data available (Bikel, Miller, Schwartz, & Weischedel, 1997). Bahl, Jelinek, and Mercer (1990) apply statistical methods to automatic speech recognition. They use smoothed n-grams in order to find the most probable string of words from a set of candidate strings. Merialdo (1994), as part of speech tagging, uses hidden Markov models to find the most probable sequence of tags in a given sequence of words. In a more complex domain, syntactic parsing using probabilistic grammars permits to find the most probable parse tree given a word sequence (Charniak, 1997). These methods have also been used for word disambiguation (Buscaldi, 2009; Stokoe, Oakes, & Tait, 2003).

2.1.2. Rule-based information extraction

Rule-based Information Extraction is a process by which structured entities are extracted from text based on rules written by human developers that capture syntactical, lexical and semantic knowledge required to identify the entities and the relationships in the domain (Feldman et al., 2002). The compositional nature of rule-based information extraction also allows rules to be expressed over previously extracted entities (Michelakis, Krishnamurthy, Haas, & Vaithyanathan, 2009).

Fig. 1 shows the schematic diagram of Rule-based IE, which has five processes and uses dictionaries or lexicons that are created by experts of a specific domain.

(a) Tokenization: Identifying the boundaries of sentences in a document and decomposing each sentence into tokens. Tokens are obtained by splitting a sentence along a predefined set of delimiters like spaces, commas, and dots. A token is typically a word, a digit, or punctuation.
(b) Part-of-speech (POS) tagging: assigning to each word a grammatical category coming from a fixed set. The set of tags includes the conventional parts of speech such as noun, verb, adjective, adverb, article, conjunct and pronoun.
(c) Semantic tagging: recognizing entities relevant to the domain. For this step it is common to use dictionaries or gazetteers that contain relevant terms and are created by experts of a specific domain.
(d) Parsing: grouping words in a sentence into prominent phrase types such as noun phrases, prepositional phrases, and verb phrases. A context-free grammar is typically used