# Chapter I
# TODE:
## An Ontology-Based Model for the Dynamic Population of Web Directories

**Sofia Stamou**
*Patras University, Greece*

**Alexandros Ntoulas**
*University of California Los Angeles (UCLA), USA*

**Dimitris Christodoulakis**
*Patras Univeristy, Greece*

## ABSTRACT

*In this chapter we study how we can organize the continuously proliferating Web content into topical categories, also known as Web directories. In this respect, we have implemented a system, named TODE that uses a **T**opical **O**ntology for **D**irectories' **E**diting. First, we describe the process for building our ontology of Web topics, which are treated in TODE as directories' topics. Then, we present how TODE interacts with the ontology in order to categorize Web pages into the ontology's topics and we experimentally study our system's efficiency in grouping Web pages thematically. We evaluate TODE's performance by comparing its resulting categorization for a number of pages to the categorization the same pages display in the Google directory as well as to the categorizations delivered for the same set of pages and topics by a Bayesian classifier. Results indicate that our model has a noticeable potential in reducing the human-effort overheads associated with populating Web directories. Furthermore, experimental results imply that the use of a rich topical ontology increases significantly classification accuracy for dynamic contents.*

## INTRODUCTION

Millions of users today access the plentiful Web content to locate information that is of interest to them. However, as the Web grows larger the task of locating relevant information within a huge network of data sources is becoming daunting. Currently, there are two predominant approaches for finding information on the Web, namely searching and browsing (Olston & Chi, 2003). In the process of searching, users visit a Web search engine (e.g., Google) and specify a query that best describes what they are looking for. During browsing, users visit a Web directory (e.g., the Yahoo! directory), which maintains the Web organized in subject hierarchies, and navigate through these hierarchies in the hope of locating the relevant information. The construction of a variety of Web directories in the last few years (such as the Yahoo! directory (http://yahoo.com), the Open Directory Project (ODP) (http://dmoz.org), the Google directory (http://dir.google.com) etc.) indicates that Web directories have gained popularity as means for locating information on the Web.

Typically, the information provided by a Web search engine is automatically collected from the Web without any human intervention. However, the construction and maintenance of a Web directory involves a staggering amount of human effort because it is necessary to assign an accurate subject to every page inside the Web directory. To illustrate the size of the effort necessary, one can simply consider the fact that Dmoz, one of the largest Web directories, relies on more than 65,000 volunteers around the world to locate and incorporate relevant information in the directory. Given a Web page, one or more volunteers need to read it and understand its subject, and then examine Dmoz's existing Web directory of more than 590,000 subjects to find the best fit for the page. Clearly, if we could help the volunteers automate their tasks we would save a lot of time for a number of people.

One way to go about automating the volunteers' tasks of categorizing pages is to consider it as a classification problem. That is, given an existing hierarchy of subjects (say the Dmoz existing hierarchy) and a number of pages, we can use one of the many machine learning techniques to build a classifier which can potentially assign a subject to every Web page. One problem with this approach however, is that in general it requires a training set. That is, in order to build an effective classifier we need to first train it on a set of pages which have already been marked with a subject from the hierarchy. Typically this is not a big inconvenience if both the collection that we need to classify and the hierarchy are static. As a matter of fact, as shown in (Chakrabarti et al., 1998a; Chen & Dumais, 2000; Huang et al., 2004; Mladenic, 1998), this approach can be quite effective. However, in a practical situation, neither the Web nor the subject hierarchies are static. For example, previous studies have shown that eight percent of new pages show up on the Web every week (Ntoulas et al., 2004) and Dmoz's subject hierarchy is undergoing a variety of changes every month[1]. Therefore, in the case of the changing Web and subject hierarchy, one would need to recreate the training set and re-train the classifier every time a change was made.

In this chapter, we present a novel approach for constructing a Web directory, which does not require a training set of pages, and therefore can cope very easily with changes on the Web or the subject hierarchy. The only input that our method requires is the subject hierarchy from a Web directory that one would like to use and the Web pages that one would like to assign to the directory. At a very high level our method proceeds as follows: first, we enrich the subject hierarchy of the Web directory by leveraging a variety of resources created by the natural language processing community and which are freely available. This process is discussed in Section 2. Then, we process the pages one by one and identify the most important terms inside every page and we link