Semi–Structured Document Classification

Ludovic Denoyer
University of Paris VI, France

Patrick Gallinari
University of Paris VI, France

INTRODUCTION

Document classification developed over the last 10 years, using techniques originating from the pattern recognition and machine-learning communities. All these methods operate on flat text representations, where word occurrences are considered independents. The recent paper by Sebastiani (2002) gives a very good survey on textual document classification. With the development of structured textual and multimedia documents and with the increasing importance of structured document formats like XML, the document nature is changing. Structured documents usually have a much richer representation than flat ones. They have a logical structure. They are often composed of heterogeneous information sources (e.g., text, image, video, metadata, etc.). Another major change with structured documents is the possibility to access document elements or fragments. The development of classifiers for structured content is a new challenge for the machine-learning and IR communities. A classifier for structured documents should be able to make use of the different content information sources present in an XML document and to classify both full documents and document parts. It should adapt easily to a variety of different sources (e.g., different document type definitions). It should be able to scale with large document collections.

BACKGROUND

Handling structured documents for different IR tasks is a new domain that recently has attracted increasing attention. Most of the work in this new area has concentrated on ad hoc retrieval. Recent Sigir workshops (2000, 2002, 2004) and journal issues (Baeza-Yates et al., 2002; Campos et. al., 2004) were dedicated to this subject. Most teams involved in this research gather around the recent initiative for the development and the evaluation of XML IR systems (INEX), which was launched in 2002. Besides this mainstream of research, some work is also developing around other generic IR problems like clustering and classification for structured documents. Clustering mainly has been dealt with in the database community, focusing on structure clustering and ignoring the document content (Termier et al., 2002; Zaki & Aggarwal, 2003). Structured document classification, the focus of this article, is discussed in greater length below.

Most papers dealing with structured documents classification propose to combine flat text classifiers operating on distinct document elements in order to classify the whole document. This has been developed mainly for the categorization of HTML pages. Yang, et al. (2002) combine three classifiers operating respectively on the textual information of a page and on titles and hyperlinks. Cline (1999) maps a structured document onto a fixed-size vector, where each structural entity (title, links, text, etc.) is encoded into a specific part of the vector. Dumais and Chen (2000) make use of the HTML tags information to select the most relevant part of each document. Chakrabarti, et al. (1998) use the information contained in neighboring documents of HTML pages. All these methods rely explicitly on the HTML tag semantic (i.e., they need to know whether tags correspond to a title, a link, a reference, etc.). They cannot adapt to more general structured categorization tasks. Most models rely on a vectorial description of the document and do not offer a natural way for dealing with document fragments. Our model is not dependent on the semantic of the tags and is able to learn which parts of a document are relevant for the classification task.

A second family of models uses more principled approaches for structured documents. Yi and Sundaresan (2000) developed a probabilistic model for tree-like document classification. This model makes use of local word frequencies specific to each node, so that it faces a very severe estimation problem for these local probabilities. Diligenti, et al. (2001) proposed the Hidden Tree Markov Model (HTMM), which is an extension of HMMs, to tree-like structures. They performed tests on the WebKB collection, showing a slight improvement over Naive Bayes (1%). Outside the field of information retrieval, some related models also have been proposed. The hierarchical HMM (Fine et al., 1998) (HHMM) is a generalization of HMMs, where hidden nodes emit sequences instead of symbols for classical HMMs. The HHMM is aimed at discovering substructures in sequences instead of processing structured data.

Generative models have been used for flat document classification and clustering for a long time. Naive Bayes (Lewis, 1998) is one of the most used text classifiers, and different extensions have been proposed (Koller & Sahami, 1997). Probabilistic models with latent variables have been used recently for text clustering, classification, or mapping by different authors. (Vinokourov & Girolami, 2001; Cai & Hofmann, 2003). Blei and Jordan (2003) describe similar models for learning the correspondence among images or image regions and image captions. All these models do not handle structured representations.

Finally, Bayesian networks have been used for the task of ad hoc retrieval, both for flat documents (Callan et al., 1992) and for structured documents (Myaeng et al., 1998; Piwowarski et al., 2002). This is different from classification, since the information need is not specified in advance. The models and problems are, therefore, different from those discussed here.

MAIN THRUST

We describe a generative model for the classification of structured documents. Each document will be modeled by a Bayesian network. Classification then will amount to performing inference in this network. The model is able to take into account the structure of the document and different types of content information. It also allows one to perform inference either on whole documents or on document parts taken in their context, which goes beyond the capabilities of classical classifier schemes. The elements we consider are defined by the logical structure of the document. They typically correspond to the different components of an XML document.

In this article, we introduce structured documents and the core Bayesian network model. We then briefly summarize some experimental results and describe possible extensions of the model.

Structured Document

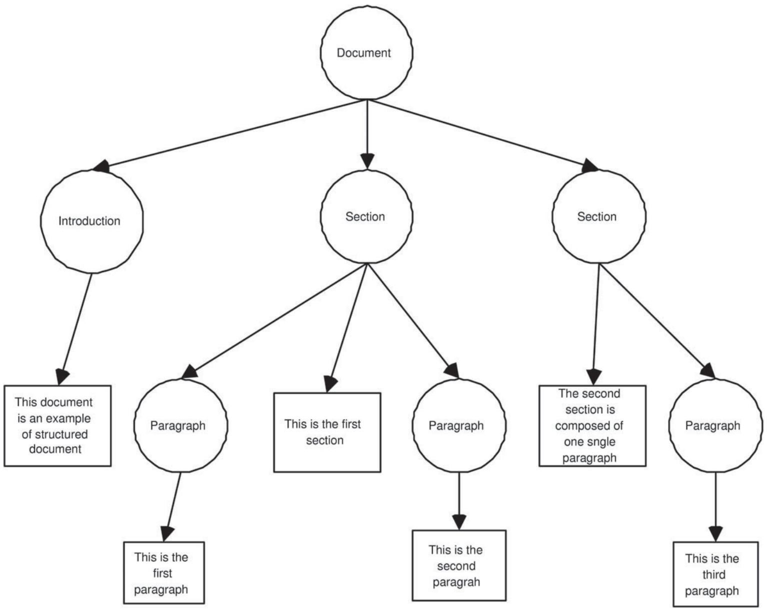
We will consider that a document is a tree, where each node represents a structural entity. This corresponds to the usual representation of XML document. A node will contain two types of information:

- A label information that represents the type of structural entity. A label could be, for example, paragraph, section, introduction, or title. Labels depend on the document's corpora; for XML documents, they are usually defined in the DTD.
- A content information. For a multimedia document, this could be text, image, or signal. For a textual document node with the label *paragraph*, the node content will be the paragraph text.

We will refer to structural and content nodes for these two types of information. Figure 1 gives an example for a simple textual document.

We will consider only textual documents here. Extensions for multimedia documents are considered in Denoyer, et al. (2004a).

Figure 1. A tree representation for a structured document composed of an introduction and two sections. Circle and square nodes are respectively structural and content nodes


Document
Introduction
Section
Section
This document is an example of structured document
Paragraph
This is the first section
Paragraph
The second section is composed of one sngle paragraph
Paragraph
This is the first paragraph
This is the second paragrah
This is the third paragraph