A similarity between probabilistic tree languages: application to XML document families

Rafael C. Carrasco, Juan Ramón Rico-Juan

Departamento de Lenguajes y Sistemas Informáticos. Universidad de Alicante. E-03071 Alicante (Spain)

Abstract

We describe a general approach to compute a similarity measure between distributions generated by probabilistic tree automata that may be used in a number of applications in the pattern recognition field. In particular, we show how this similarity can be computed for families of structured (XML) documents can be computed. In such case, the use of regular expressions to specify the right part of the expansion rules adds some complexity to the task.

Key words: Distance between tree languages; similarity of structured documents.

1 Introduction

Trees become an adequate representation of data in those tasks where the data components keep a hierarchical relation. For instance, they are a suitable representation of syntactic parses, graphic patterns, segmented images or structured documents. Sometimes, comparing families of trees is of interest. For example, a *document type definition* (DTD) for XML documents defines by means of a EBNF context-free grammar [1] a tree language where the structural tags are represented by node labels. Even if all valid documents must comply with the DTD, it is possible that some subclasses of documents show different typical patterns, that is, they can be modeled by different probability distributions over the language of valid structures specified by the DTD.

[1] Work supported by the Comisión Interministerial de Ciencia y Tecnología through grant TIC2000-1599-C02 and by the Oficina de Ciència y Tecnologia de la Generalitat Valenciana under grant GV01-316.

Often, tree languages are infinite and an efficient measure to compare them is needed. For this purpose, the Kullback-Leibler divergence [2] is not suitable whenever a single element has a null probability only in one of the languages and then, the result becomes infinite. Furthermore, we are also interested in a similarity measure suitable when regular expressions can be used on the right hand side of the production rules (as in DTDs and, more generally, in EBNF specifications).

As noted by Lyngsø, Pedersen and Nielsen [4], both the quadratic distance and a cosine-type similarity can be computed in the case of Hidden Markov Models (HMM) with simple cycles. However, it is not difficult to generalize this result, following [3], to compute the similarity between general stochastic finite-state automata or HMMs that operate on strings. Here, we will show how to compute this measure for tree languages modeled by probabilistic tree automata described by means of a EBNF context-free grammar.

2 Generative models

Structured documents can be regarded as trees whose tags belong to an alphabet Σ and follow a set R of rules. Usually, rules of the type $\texttt{doc} \rightarrow ((\texttt{head}, \texttt{para}*)|\texttt{para}+)$, describe the valid use of structural tags such as `doc`, `head` or `para`. However, more general specifications allow the tags not to define completely the possible element expansions. For instance, RELAX grammars [5] differentiate between *roles* (structural tags) and *labels* (in the following called *states*). Therefore, elements with the same tag may indeed belong to different states and, then, the applicable expansion rules differ. This approach is based on the theory of tree automata and suggests the following definition.

An *extended probabilistic tree automaton* $A = (Q, \Sigma, R, \rho, P)$ consists of a finite set of *states* Q; a finite set of tags or *alphabet* Σ; a set R of expansion *rules* of the form $q \rightarrow f(\alpha_q^f)$ where $q \in Q$, $f \in \Sigma$ and α_q^f is a regular expression over Q; a function ρ that gives the probability $\rho(q)$ that a tree root belongs to state q; and a probability function P that for each state q and for each expansion $f(q_1 \cdots q_m)$ such that $q_1 \cdots q_m$ matches α_q^f gives the conditional probability $P(f(q_1 \cdots q_m)|q)$.

The probability that the automaton A assigns to a tree t in T_Σ, the class of trees with labels in Σ, is

$$\sum_{q \in Q} \rho(q)\ p(t|q) \tag{1}$$

where $p(t|q)$ is recursively defined for every $q \in Q$ and for every tree $f(t_1, \ldots, t_m)$

made of a root node with label $f \in \Sigma$ and $m > 0$ subtrees $t_1, \ldots, t_m$ as

$$p(f(t_1, \ldots, t_m)|q) = \sum_{(q_1,\ldots,q_m)\in Q^m} p(f(q_1 \cdots q_m)|q)\; p(t_1|q_1) \cdots p(t_m|q_m) \qquad (2)$$

and as $p(f()|q) = P(f()|q)$ for leaves. We will assume that the probability of the q-expansion $f(q_1 \cdots q_m)$ is given by that of the string $fq_1 \cdots q_m\#$ (using $\#$ to mark the end-of-string) in a HMM $M_q = (S_q, V, a_q, b_q, \pi_q)$ with states S_q, observable alphabet $V = \Sigma \cup Q \cup \{\#\}$, state transition probabilities $a_q(i, j)$, observation symbol probabilities $b_q(i, r)$ and initial state distribution $\pi_q(i)$.

3 Probabilistic measures

The probability that two models A and A' generate the same tree

$$\mathcal{C}(A, A') = \sum_{t\in T_\Sigma} p(t)p'(t) \qquad (3)$$

allows one to define a similarity measure [4] as

$$\cos(A, A') = \mathcal{C}(A, A')/\sqrt{\mathcal{C}(A, A)\; \mathcal{C}(A', A')} \qquad (4)$$

If, for every $q \in Q$ and $r \in Q'$ we define $\eta_{qr} = \sum_t \pi(t|q)\pi'(t|r)$, then

$$\mathcal{C}(A, A') = \sum_{q\in Q} \sum_{r\in Q'} \eta_{qr}\rho(q)\rho'(r). \qquad (5)$$

If we now denote with $\eta_{qr}^{[ij]}$ the following contribution to η_{qr}:

$$\eta_{qr}^{[ij]} = \sum_m \sum_{(t_1,\ldots t_m)\in T_\Sigma^m} \sum_{(q_1,\ldots q_m)\in Q^m} \sum_{(r_1,\ldots r_m)\in Q^m} \alpha_q(i, q_1 \cdots q_m\#)\; p(t_1|q_1) \cdots p(t_m|q_m)$$
$$\alpha'_r(j, r_1 \cdots r_m\#)\; p(t_1|r_1) \cdots p(t_m|r_m)$$

where α and α' represent classical forward probabilities in the HMMs, it is not difficult to realize that

$$\eta_{qr} = \sum_{f\in\Sigma} \sum_{k\in S_q} \sum_{l\in S'_r} \pi_q(k)\; \pi'_r(l)\; b_q(k, f)\; b_r(l, f)\; \eta_{qr}^{[kl]} \qquad (6)$$

Finally, the coefficients $\eta_{qr}^{[ij]}$ can be simply computed by solving (e.g., iteratively) a linear system of equations:

$$\eta_{qr}^{[ij]} = b_q(i, \#)\; {b'}_r(j, \#)$$
$$+ \sum_{s\in Q} \sum_{t\in Q'} \sum_{k\in S_q} \sum_{l\in S'_r} a_q(i, k)\; {a'}_r(j, l)\; b_q(k, s)\; b'_r(l, t)\; \eta_{st}\; \eta_{qr}^{[kl]} \qquad (7)$$

4 Preliminary results and conclusion

The model has been used to compute similarities between XML document sets of different authors in the Miguel de Cervantes Digital Library [2]. The HMMs states and probabilities were extracted from the Glushkov automata [1] of the regular expressions in the DTD and from the local element frequencies in each collection. The results reflect that some author works deviate from the average

	A1	A2	A3	A4
A2	0.015			
A3	10^{-7}	0.548		
A4	0.043	0.311	0.059	
All	0.004	0.871	0.798	0.205

Fig. 1. Similarity between different collections: Mariano J. de Larra (A1), Clarín (A2), Concepción Arenal (A3), Juan Valera (A4) and the whole XML library (All)

structure: the M.J. de Larra collection mainly consists of journal articles, C. Arenal's are political essays while Clarín's and Valera's collections are mainly novels, the most common genre in the library. A proof of equation (7) and further results will be shown in a forthcoming longer version of the paper.

References

[1] Anne Brüggemann-Klein and Derick Wood. One-unambiguous regular languages. *Information and Computation*, 142(2):182–206, 1998.

[2] Jorge Calera-Rubio and Rafael C. Carrasco. Computing the relative entropy between regular tree languages. *Information Processing Letters*, 68(6):283–289, 1998.

[3] Rafael C. Carrasco. Accurate computation of the relative entropy between stochastic regular grammars. *RAIRO (Theoretical Informatics and Applications)*, 31(5):437–444, 1997.

[4] Rune B. Lyngsø, Christian N. S. Pedersen, and Henrik Nielsen. Metrics and similarity measures for hidden Markov models. In T. Lengauer et al., editor, *Proc. of the 7th Int. Conference on Intelligent Systems for Molecular Biology (ISMB-99)*, pages 178–186, Menlo Park, CA, 1999. AAAI Press.

[5] Murata Makoto. Document description and processing languages. regular language description for XML (RELAX). Technical report, ISO/IEC DTR 22250-1, 2001.

[2] `http://cervantesvirtual.com`