Exam 2016 – Web Data Management

S. Abiteboul & P. Senellart

1 Simple Proabilistic XML (14 points)

In this exercise, we define an *XML document* (or *document* for short) as an unordered, unranked, labeled tree. A *simple probabilistic XML document* (or *sp-document* for short) is a document where every node n except for the root is annotated with a probability $\pi(n) \in (0; 1]$.

An sp-document D defines a random process as follows:

- independently for every node n of D, decide whether to keep or discard n with probability $\pi(n)$;
- remove every discarded node n from the tree along with all its descendants;
- remove all probability annotations.

The result is a probability distribution $[\![D]\!]$ over documents, called the *semantics* of D. The support of $[\![D]\!]$ is called the *set of possible documents* of D.

A *tree-pattern query* q is an XPath path expression where all predicates are relative path expressions (that test for the existence of a path) and such that only the `child` and `descendant` axes are used. For a document d, we denote by $q(d)$ the Boolean evaluation of q on d.

1. (1 point) Give two significant differences between XML documents as defined here and XML documents as defined in the XML standard.
2. (2 points) Give an example of an sp-document D with exactly 6 possible documents, along with a full description of $[\![D]\!]$.
3. (1 point) Give an example of a tree-pattern query with exactly two predicates.
4. (2 points) Propose a simple algorithm to evaluate $q([\![D]\!])$, the probability that $q(d)$ is true for $d \in [\![D]\!]$. What is the complexity of this algorithm?
5. (2 points) Show that there exist finite sets of possible documents whose roots have the same label and that are not the support of the semantics of an sp-document.
6. (2 points) We consider now the non-Boolean evaluation of tree-pattern queries: the result of a query is the minimal subtree with same root containing all nodes matched by subparts of the query. Show that sp-documents are not a strong representation system[1] for non-Boolean tree-pattern queries.
7. (2 points) Propose a strong representation systems for probability distributions over documents for non-Boolean tree-pattern queries. This representation system should be exponentially more compact than the enumeration of possible documents, for infinitely many sets of possible documents.
8. (2 points) Show that there is no representation systems for probability distributions over documents for non-Boolean tree-pattern queries such that representation sizes are always exponentially more compact than the enumeration of possible documents.

2 Essay (6 points)

Take a large-scale Web-based service of your choice (e.g., Google Search, Bing Maps, Wikipedia, Reddit, Amazon, Facebook, Twitter, etc.) Give some order of magnitudes of the data and queries this service needs to store and process (1 point). Explain what are the most critical features that such a service needs to support (1 point). Propose two possible architecture choices to implement such a service at scale and describe them in detail (2 points). Discuss the advantages and disadvantages of these architectures (2 points). The entire essay should not take more than two pages.

[1] A representation system for probability distributions over objects is *strong* for a query class if the probability distributions of query answers over these objects can be represented in the same representation system.