Proposition de recherche doctorale

Uncertainty over Intensional Data

Mots clés :
● Directeur de thèse : PIERRE SENELLART
● Co-encadrant(s) :
● Unité de recherche : Laboratoire Traitement et Communication de l'Information
● Ecole doctorale : Ecole Doctorale Informatique, Télécommunications, Électronique de Paris
● Domaine scientifique principal: Divers

Résumé du projet de recherche (Langue 1)
In contrast with traditional data management applications, a wealth of useful data available today is uncertain and limited by access restrictions. For instance, though originally intended for human consumption, the World Wide Web is currently the most comprehensive information source accessible to automated agents. A major challenge is to structure this information and answer complex queries over it. To do so, many types of data sources must be integrated: semantic Web sources such as SPARQL endpoints which provide structured data; hidden Web sources out of which information can be extracted by crawling and wrapper induction; semi-structured HTML documents and natural language text on which information extraction can be applied; and possibly human sources in the context of crowdsourcing. However, information on the Web is {uncertain}: first, the data itself can be biased, wrong, or annotated with confidence values; second, uncertainty can accumulate over the various steps needed to extract the data in a structured form. Besides, information on the Web is {intensional}: the Web does not reside on one machine, but is distributed and must be accessed piece by piece by querying a wide range of separate independent services. For this reason, the interplay between uncertainty and intensionality is at the heart of the management of structured data obtained from the Web. Uncertain data can be seen mathematically as a set of possible states of the data (i.e., the possible worlds), which can be equipped with a probability distribution. However, because there can be an infinite number of possible worlds, a finite implicit representation of the uncertain data is needed to perform computation on it. There are numerous formalisms for such representations, depending on the kind of data being studied. Probabilistic relational databases store uncertain relational data and are designed to answer relational queries efficiently. For tree-like data such as XML, numerous probabilistic frameworks exist. On a more fundamental level, the usual representation of an uncertain string is a weighted finite-state automaton for the string, which links the management of uncertainty to the theory of formal languages. The different ways to represent uncertain data can be compared according to their expressiveness (which probability distributions can be represented by a given model and how are the representations computed?), their size (how large is the representation of a distribution and how far can a representation be compressed or approximated?), and their efficiency (what is the complexity of computing such a model from examples, or of answering queries over such models?). The performance of the various models over these criteria and the relations between those models are not yet adequately understood. Unlike in traditional data management settings, data from the Web must often be examined in an intensional manner: the data is not really present on the computing device but can be provided on-demand by separate services. Hard constraints such as access patterns, or soft constraints such as bandwidth limitations, may be imposed when accessing the services. In this setting, several important problems arise. First, the intensional nature of the Web makes it difficult to determine if the set of available access patterns is sufficient to evaluate a given query. The decidability and complexity of problems such as the immediate or long-term relevance of an access for a certain query have been studied and is closely linked to the problem of query containment for various query languages. Second, even if a query is feasible, it may be hard to find an execution plan which does not impose unreasonable load on the services. Because crawling the entire Web to answer a query is not realistic, we must limit the number of accesses we perform. In extreme cases such as crowdsourcing in which some services are provided by humans, the cost of each individual access may be very high. Uncertainty can be used as a way to manage intensionality: data which has been accessed can be uncertain, but data which has not yet been accessed is even more uncertain, and we can see queries as a way to spend resources to reduce uncertainty. Uncertainty can also be used as a way to manage overlapping, conflicting or incomplete information from independent intensional sources. Ideally, our estimation of the uncertainty of data should be a combination of the existing probability distribution of the data, the combined confidence of the various extraction steps, the proportion in which involved sources agree, and a tradeoff between harvesting more information and limiting the resources used by a query.

Résumé du projet de recherche (Langue 2)
Possible goals of this PhD include: -* Understanding connections between various representations of uncertainty and differences in their expressiveness and compactness. -* Characterizing subsets of probabilistic data which achieve a good balance between expressiveness and complexity. -* Developing methods for compact representation of probabilistic data, possibly up to approximations. -* Achieving a broader understanding of the impact of access restrictions on intensional data, from the angle of decidability, complexity, and query planning. -* Proposing models for the integration of heterogeneous data sources such as SPARQL endpoints, deep Web crawling, and information extraction results, as well as the results of crowdsourcing, which reflect both their uncertain and intensional aspects. -* Prototyping novel results to validate their practical effectiveness.

Informations complémentaires (Langue 1)
Collaborations are planned with the Information Systems group at the University of Oxford (Prof. Michael Benedikt) and with the Database group at Tel Aviv University (Prof. Tova Milo).