Draft preview -- not a final published version

Generated Sun Mar 2 12:04:08 CET 2014

UnSAID: Uncertainty and Structure in the Access to Intensional Data

Antoine Amarilli Pierre Senellart
Institut Mines–Télécom; Télécom ParisTech; CNRS LTCI; Paris, France
firstname.lastname@telecom-paristech.fr

ABSTRACT

To answer user queries on Web data, it is necessary to crawl, extract, enrich, and process available information. The traditional *extensional* approach is to perform those steps one after the other, but it has many drawbacks. The choice of information that we retrieve and process must be guided by the query, because retrieving all the information is not feasible; the information cannot be maintained locally because it may become obsolete rapidly; it cannot be trusted blindly, as it may come from untrustworthy sources; it must be stored in a way which accounts for its heterogeneous structure (Web pages, relational facts, textual content, etc.). In this paper, we present UnSAID, our vision of a framework which addresses simultaneously the three main challenges faced by the extensional approach: *intensionality*, the need to access data selectively and take into account the cost of individual accesses; *uncertainty*, the need to reason on partial and inexact views of the world; and *structure*, the need to deal with data in various heterogeneous forms.

1. INTRODUCTION

Publicly available data, information, knowledge is abundant: the World Wide Web contains trillions of pages on an amazingly diverse collection of topics; hundreds of thousands of deep Web databases, accessible through Web forms, are also available; a social networking site such as Twitter sees hundreds of millions of new (public) messages posted *each day*; the open linked data now contains hundreds of knowledge bases covering tens of billions of semantic facts in the form of RDF triples; complex tools in areas such as information extraction, data mining, or natural language processing (NLP) are readily available to enrich existing data with even more information; rules mined from data, or machine learning models, can be used to make predictions; and when the data is not there and cannot be predicted, or when it is not easy to process automatically, it is always possible to resort to crowdsourcing platforms such as Amazon Mechanical Turk to collect or annotate data.

Yet, the availability of data does not mean that it can be leveraged easily to satisfy a user's needs. We call *knowledge acquisition needs* the demands which can be phrased by users: they may correspond to precise queries, such as "does a certain fact hold?", or more vague requests, such as "find all relevant information about a certain topic". Many challenges need to be addressed to satisfy such needs: the available data sources are numerous and heterogeneous, accessing data carries a certain cost, and some of the available data may be imprecise or incorrect.

As a first example of the approach, consider the application of *mobility in smart cities*, i.e., a system integrating information about transportation options, travel habits, traffic, etc., in and around a city. All resources mentioned in the previous paragraph can be used to collect and enrich data related to this application: the Web, deep Web sources, social networking sites, the Semantic Web, annotators and wrapper induction systems, crowdsourcing platforms, etc. Moreover, in such a setting, domain-specific resources, not necessarily public, contribute to the available data: street cameras, red light sensors, air pollution monitoring systems, etc.

Users of the system, namely, transport engineers, ordinary citizens, etc., may have many kinds of knowledge acquisition needs. They can be simple queries expressed in a classical query language (e.g., "How many cars went through this road during that day?" or "What is the optimal way to go from this place to that place at a given time of day?"), certain patterns to mine from the data ("Find an association rule of the form $X \Rightarrow Y$ that holds among people commuting to this district."), or higher-level business intelligence queries ("Find anything interesting about the use of the local bike rental system in the past week.").

As a second example, consider the problem of *personal information management*, namely, integrating user data across services that manage the user's emails, calendar, social network, travel information, etc. To answer a knowledge acquisition need such as "find the people I need to warn about my upcoming trips", the system would have to orchestrate queries to the various services: extract the trips, identify the meetings that conflict with them, and determine their likely participants.

As a third example, consider *socially-driven Web archives* [26]: their goal is to build semantically annotated Web archives on specific topics or events (investment for growth in Europe, the 2014 Winter Olympics, etc.), guiding the process with clues from the social Web as to which documents are relevant. These archives can then be semantically queried by journalists today or historians tomorrow, e.g., to retrieve all resources mentioning a given person. The construction of these archives relies on Web crawling, deep Web harvesting, access to social networking sites such as Twitter or YouTube via their APIs, use of tools for information extraction, named entity recognition, opinion mining, etc.

The *unsaid* is, according to Wikipedia, "what is hidden and/or implied". We claim, as illustrated by these three scenarios, that most of the data from the Web and other sources that is useful to solve a user's specific knowledge need is, similarly, hidden and not explicitly present, but needs to be crawled, extracted, annotated, by performing costly accesses to sources. Our vision, UnSAID, is that of a system to answer a user's needs by taking into account the heterogeneity of content, the cost in accessing it, and its uncertainty,

In this vision paper, we first briefly describe the high-level *extensional* approach which would currently be used to tackle the scenarios we described, and outline how we think they fall short of solving them (Section 2). The UnSAID approach is then presented

Draft preview -- not a final published version

Generated Sun Mar 2 12:04:08 CET 2014

in Section 3 and illustrated on our example use cases in Section 4. We then discuss in Section 5 how UnSAID relates to the state of the art in Web data management before concluding in Section 6.

2. EXTENSIONAL APPROACH

Let us review the traditional approach, which we dub the *extensional* approach, to answer such knowledge acquisition needs. First, the system would collect all available data and dump it in a data warehouse, dealing with schema mapping issues along the way; or it would collect all data from each source independently, keep it in its original form, and use a mediator system [29] to have a global view of the collection. The system would then enrich, annotate, and curate the extracted data using automatic tools, expert feedback, and crowdsourcing. Last, it would run whatever query or mining operation is needed on this data.

However, this approach ignores the fact that accessing the data has a *cost*, in terms of HTTP requests and bandwidth, rate policy limitations to access social networking APIs, budget to pay crowdworkers, computation time needed to run NLP tools or reasoners, etc. Hence, the extensional approach does not scale: there is simply too much data available to collect, too many potentially useful services to run, etc. One must identify *a priori* a subset of the data that can reasonably be extracted; however, it is impossible to determine ahead of time what data will or will not be relevant to the present and future user needs. In addition, storing extracted data locally means that it has to be refreshed whenever the data sources are updated, which may be even more costly.

In line with existing terminology,[1] we call such data that is accessible but only after making an access, after paying some cost, *intensional data*. Hence, the first drawback of the traditional approach is that it does not account for the *intensionality* of data.

The second shortcoming of the traditional approach is that it will usually not account for the *uncertain* character of most of the data that is collected or produced. Web sources might not be trustworthy, automated annotators may be unreliable, rules may be uncertain, crowdworkers may provide wrong answers, and the overall set of facts is in general contradictory.

When uncertainty is not outright ignored, it is usually only considered locally: for instance, when extracting information from a Web page, a probabilistic model may be used to represent all possible annotations, but then only the best answer will be returned, and the uncertainty is forgotten in subsequent processing. However, if the entire process only retains the most likely answer at each step, the end result may not be the most likely overall. Besides, if no information is kept about the inherent uncertainty of data items, or their *provenance*, the user will be unable to judge of the potential quality of the overall result. Indeed, for most open-ended knowledge acquisition needs on uncurated data sources like the Web, information is essentially worthless unless its source is taken into consideration.

The third drawback of the traditional approach is that it will usually force all the extracted data in flat relational databases, or store them separately in their original formats without trying to integrate them. Yet, in addition to relational tables, there is a variety of other *structures* in which data may reside: (i) semi-structured tree-like content, such as Web pages or the result of wrapper extraction systems on top of Web pages; (ii) graph data, in particular social networking data, semantic graphs, or traffic networks such as those of OpenStreetMap; (iii) spatio-temporal data, sometimes (like in Twitter or sensor networks) in the form of time-ordered streams; (iv) complex views with aggregation (e.g., a source may list the number of traffic accidents per district of a city over a year, which is an aggregation of an underlying database of traffic accidents). Fitting widely heterogeneous data into a uniform schema does not respect its inherent structure. For instance, some operations, such as descendant queries, are natural on tree-like data but harder to express in the relational setting; whereas relational operations such as joins may make little sense, e.g., for graph-shaped data.

Those three shortcomings of traditional knowledge acquisition approaches are not independent; rather, they interact in a tightly coupled fashion. For example, if we use a probabilistic modeling of uncertainty, we need to represent, manage, query, probability distributions on structured objects, so that the representation of uncertainty depends on the structure that we use. Likewise, the kinds of intensional accesses which we may perform depend on the structure of the data that we consider. Last, the intensional accesses which we perform will depend on our representation of uncertainty, as this representation may be used, e.g., to represent our prediction on the results of accesses which have not been perfomed yet.

3. UNSAID

Having described the challenges that must be faced to solve knowledge acquisition needs, we now describe our general approach to address them. We call the approach *UnSAID*, standing for *Uncertainty and Structure in the Access to Intensional Data*.

We choose to model uncertainty in the data as probability distributions over the state of the world. Other forms of uncertainty models exist (in particular, fuzzy sets and Dempster-Shafer theory), but probability theory offers many advantages: a clean mathematical framework that makes it possible to perform precise computations, a set of well-understood and tractable sampling-based techniques, and wide existing use in the form of probabilistic frameworks for information extraction or NLP that naturally produce probabilities [19].

At every point in time, the system has a *partial* and *uncertain* view of the whole data of interest. We thus see its current knowledge as a probabilistic distribution over all possible worlds, and this knowledge evolves as the system discovers new things about the world. As we cannot represent explicitly a probability distribution on the (generally infinite) set of possible states of the world, we can use *probabilistic representations* [17, 27] to represent our current knowledge concisely.

We view available resources as *services* that take as input some data items and return, at a cost, a potentially uncertain set of new data items. Thus, a service can be used to refine our existing knowledge. While some services require an input (e.g., retrieve the emails of a given user), others do not (e.g., data readily available from the Web, that only requires the cost of an HTTP request). Some services may be associated with *a priori* information on the data they may return, in the form of a probability distribution over possible responses. Here, there are two sources of uncertainty: uncertainty about what an access may return before performing it, and uncertainty in the results of service application themselves.

This representation makes it possible to reason about whether an access is worth the cost, before we perform it. To determine the value of an access, we must estimate how much we expect it to reduce our uncertainty about the current state of the world, weighting it relative to the knowledge acquisition need posed by the user. When considering the cost of the access, we must evaluate it along multiple heterogeneous dimensions: financial cost, CPU time, bandwidth limits, policy constraints, etc. For this reason, we cannot aggregate cost to a unidimensional variable, contrarily to what happens in traditional query optimization. This implies that multi-objective optimization may be needed.

Given a knowledge need (e.g., a query, some mining task) the problem now becomes: how to answer this need with some proba-

[1] In Datalog, *intensional* data is data not initially present but obtained through rule applications.

Draft preview -- not a final published version

bilistic guarantee on the quality of the result (say, a 95% chance that our answer will be within 5% of the exact result)[2] while minimizing the cost (the cost should take into account service calls but also query evaluation on the current knowledge).

We are thus looking for a *knowledge acquisition plan*, a plan (as in classical database optimizer plans [13]) whose operators are service calls as well as more traditional data manipulation steps, and whose objective is to answer the user's knowledge need. Such a plan may involve recursion (as it may be necessary to repeatedly and unboundedly call a given service, e.g., until there are no more results to return), and must be dynamic (as opposed to a static plan determined in advance) so as to adapt to service results. The problem is thus to determine *what is the next best thing to do to solve a knowledge need* at every point in the evaluation. Actually, the optimal plan at every point in time may also be modeled as a probability distribution over all plans, each plan being associated with its probability of being optimal given the current knowledge of the world. This goes far beyond traditional query optimization.

4. USE CASES

We now illustrate how UnSAID would proceed in the example scenarios described in the introduction. For the first application, imagine that a civil engineer wants to know the total traffic through a given road on an arbitrary day, in order to plan a renovation of that road. There are many ways to accomplish such a task: use a computer vision program to analyze the street camera feeds and identify each passing vehicle; ask crowdworkers to perform the same analysis; do this only a fraction of the day, and extrapolate the results; use traffic data from Bing Maps API, correlated with external data about road characteristics; ask expert traffic specialists to survey the road; etc. Each of these (and each combination of these) has a cost (in terms of manpower, budget, processing time, bandwidth) and a precision (both as a prior and as a posterior after using the services). UnSAID would determine an optimal solution given an approximation tolerance. This example was fairly simple, but realize that determining the traffic on a road may only be one component of more complex information needs, such as route planning.

In personal information management, recall the example knowledge acquisition need of finding the people impacted by upcoming trips. The trips may be retrieved directly from the user's travel information Web site, if one exists, but also maybe from their calendar, or from their email (hotel bookings, airline confirmations, etc.). This information has to be retrieved in an intensional manner: for instance, costly information extraction tools should only be run on emails which appear to be relevant, perhaps identified by a less costly cursory analysis. Impacted people can be found by examining events in the user's calendar and determining who is likely to attend the event, maybe based on email exchanges or participant lists for former events. Of course, uncertainty has to be maintained along the whole process, and provenance information is useful to explain query results to the user (e.g., indicate the meetings and trips that are relevant to a certain person of the output).

Last, for socially-driven Web archives, UnSAID would proceed in a manner reminiscent of *focused crawling*: from the currently known resources, retrieve the ones which we estimate are the most likely to be relevant. However, it should take into account the heterogeneity of sources and of cost (for instance, the various API rate limits, which differ depending on the provider). Furthermore, it should allow for complex queries ("find all positive statements from people who are currently attending the Olympics"), which complicates the choice of which access to perform next, and making it necessary to use, e.g., named entity recognition and sentiment analysis, costly processes which should only be run on promising resources.

[2]In other words, this is a *probably approximately correct* algorithm.

5. RELATED WORK

Probabilistic databases. Probabilistic databases, whether relational [27] or XML [17], are compact representation systems for probabilistic distributions over regular (relational or XML) databases. Research on probabilistic databases has focused on their expressiveness, on the efficiency of query evaluation (complexity bounds and efficient algorithms for the tractable cases), exact or approximate, and on building practical systems that can be used to manage probabilistic data. The literature has mostly looked at *static* scenarios; only few works consider updates [16] though there are of prominent importance in the UnSAID setting where our knowledge of the world is constantly evolving. Moreover, with few exceptions [8], the world is assumed to be closed: possible worlds are subsets of a given, finite, database. In UnSAID, we are in an open world and possible worlds may be infinitely numerous.

Reinforcement learning. Determining *the next best thing to do* under an evolving knowledge of the real world is the focus of reinforcement learning [28]. Its objective is to maximize the cumulative reward that is obtained when performing some actions, each action leading to an individual reward and to a new state, usually in a stochastic manner. Markov decision processes [25], aka MDPs, are a common model for reinforcement learning scenarios, where each action leads probabilistically to a new state and a given reward; while the underlying probability distributions are unknown, they can be learned as actions are tried. This implies an inherent tradeoff between exploration (trying out new actions yielding to new states and to potentially high rewards) and exploitation (performing actions already known to yield high rewards), a compromise that has been studied in depth in, e.g., the stateless model of multi-armed bandits [4]. There are many challenges to using MDPs in the UnSAID context, however: (i) our state space is typically huge, representing all possible partial knowledge of the world; (ii) states have complex structures, namely that of the data; (iii) rewards are typically delayed, as queries may only be answerable after a long sequence of accesses. Note that, because of data uncertainty, there is only *partial observability* of the current state, meaning considering partially observable MDPs.

Adaptive query evaluation. To find the best plan to evaluate queries on the data, traditional database management systems [13] perform query optimization using statistics gathered on the original data, which may lead to sub-optimal plans because these statistics only form a partial view of the data. Adaptive query evaluation [11] covers a wide range of techniques used to adapt the query plan to the actual performance of the query while it is evaluated: this may consist, for example, in adding query optimization operators into the query plan, or in trying out several plans in parallel until the best one is determined, on subsets of the data. To the best of our knowledge, adaptive query evaluation does not deal, however, with some aspects of the UnSAID setting, such as uncertainty or recursive query plans.

Query answering over intensional data. There are two major challenges in query answering over data accessible through services. First, as the data is costly to access, operators that need access to the data must be delayed as much a possible, which can result in lazy evaluation strategies, as in Active XML [1]. Second, services commonly limit the user in the kind of accesses that can be performed; typically, some input must be provided to the service for a corresponding output to be generated. Under such access constraints, it is typically intractable to answer even the basic question of determining whether a given service is relevant [6, 7].

Generated Sun Mar 2 12:04:08 CET 2014

Draft preview -- not a final published version

Reasoning under incompleteness and uncertainty. Regarding reasoning on incomplete data, existing work has studied the important question of *query answering* under logical constraints, namely, finding out if a fact is a logical consequence of a set of known facts and rules [10]; for instance in the context of ontology-based data access. Approaches such as backwards chaining attempt to deal with such problems in an *intensional* way, without materializing all consequences of the rules. However, in the context of *uncertain rules*, such as those obtained by a real-world knowledge base [12], state-of-the-art approaches [14] proceed by reduction to certain rules. This leaves room for a principled study of the problem accounting for, e.g., dependencies between derived facts, and more general rule languages such as existential rules.

Focused and deep-Web crawling. Focused crawlers provide an effective way to balance cost, coverage, and quality of data collection from the Web by selectively crawling pages relevant to a set of topics, for instance defined by keywords [21]. Here also, it is required to determine what is the next best URL to crawl, but the objective is limited to topic matching, on graph structures.

Some crawlers attempt to explore the *deep Web* of content hidden behind Web forms [9] by filling them automatically and retrieving results from the underlying databases. Most of them siphon the entire database in an extensional manner [5, 18] but an intensional approach is possible [22]. However, current deep Web crawlers do not support complex knowledge needs or data uncertainty.

Crowdsourcing. The field of crowdsourcing, and *crowd data sourcing* in particular, investigates how data may be obtained through queries to human operators. When doing so, one must consider the problem of achieving sufficient certainty on the answers [3, 23] as well as that of choosing the next best access to perform [2, 24]. However, works that account for uncertainty are restricted to simple data structures and knowledge acquisition needs, such as independent Boolean or multivalued questions. By contrast, crowdsourcing works which handle more complex requests [20] are not really suited to the UnSAID setting because do not follow the adaptive approach of choosing questions on-line based on previous answers.

6. CONCLUSION

In this paper, we have outlined the shortcomings of the traditional approach to answer knowledge acquisition needs, following three coupled dimensions of the problem: intensionality, uncertainty, and structure. We have outlined our vision for a new approach, UnSAID, which addresses these three challenges, and illustrated it on examples. The UnSAID objective is ambitious and can be approached from several directions. We have, in previous works, attacked this objective from the point of view of focused crawling for social networks [15] as well as crowdsourcing [2, 3]. We now intend to achieve the full UnSAID vision; our first steps are to investigate the applicability of Markov decision processes to intensional accesses, or the possible methods to reason on the intensional consequences of uncertain rules.

7. REFERENCES

[1] S. Abiteboul, O. Benjelloun, B. Cautis, I. Manolescu, T. Milo, and N. Preda. Lazy query evaluation for Active XML. In *SIGMOD*, 2004.

[2] A. Amarilli, Y. Amsterdamer, and T. Milo. On the complexity of mining itemsets from the crowd using taxonomies. In *ICDT*, 2014.

[3] Y. Amsterdamer, Y. Grossman, T. Milo, and P. Senellart. Crowd mining. In *SIGMOD*, 2013.

[4] J.-Y. Audibert, R. Munos, and C. Szepesvári. Exploration-exploitation tradeoff using variance estimates in multi-armed bandits. *Theor. Comput. Sci.*, 410(19), 2009.

[5] L. Barbosa and J. Freire. Siphoning hidden-web data through keyword-based interfaces. In *SBBD*, 2004.

[6] M. Benedikt, P. Bourhis, and P. Senellart. Monadic Datalog containment. In *ICALP*, 2012.

[7] M. Benedikt, G. Gottlob, and P. Senellart. Determining relevance of accesses at runtime. In *PODS*, 2011.

[8] M. Benedikt, E. Kharlamov, D. Olteanu, and P. Senellart. Probabilistic XML via Markov chains. *PVLDB*, 3(1), 2010.

[9] BrightPlanet. The Deep Web: Surfacing hidden value, 2000.

[10] A. Calì, G. Gottlob, G. Orsi, and A. Pieris. On the interaction of existential rules and equality constraints in ontology querying. In *Correct Reasoning*, 2012.

[11] A. Deshpande, Z. G. Ives, and V. Raman. Adaptive query processing. *Foundations and Trends in Databases*, 1(1), 2007.

[12] L. A. Galárraga, C. Teflioudi, K. Hose, and F. M. Suchanek. AMIE: association rule mining under incomplete evidence in ontological knowledge bases. In *WWW*, 2013.

[13] H. Garcia-Molina, J. D. Ullman, and J. Widom. *Database systems - the complete book (2. ed.)*. Pearson Education, 2009.

[14] G. Gottlob, T. Lukasiewicz, M. V. Martinez, and G. I. Simari. Query answering under probabilistic uncertainty in Datalog+/- ontologies. *Ann. Math. Artif. Intell.*, 69(1), 2013.

[15] G. Gouriten, S. Maniu, and P. Senellart. Exploration adaptative de graphes sous contrainte de budget. In *BDA*, 2013. Conference without formal proceedings.

[16] E. Kharlamov, W. Nutt, and P. Senellart. Updating probabilistic XML. In *Updates in XML*, 2010.

[17] B. Kimelfeld and P. Senellart. Probabilistic XML: Models and complexity. In *Advances in Probabilistic Databases for Uncertain Information Management*. Springer, 2013.

[18] J. Madhavan, A. Y. Halevy, S. Cohen, X. L. Dong, S. R. Jeffery, D. Ko, and C. Yu. Structured data meets the web: A few observations. *IEEE Data Eng. Bull.*, 29(4), 2006.

[19] C. Manning and H. Schütze. *Foundations of Statistical Natural Language Processing*. MIT Press, 1999.

[20] A. Marcus, E. Wu, D. R. Karger, S. Madden, and R. C. Miller. Human-powered sorts and joins. *PVLDB*, 5(1), 2011.

[21] F. Menczer, G. Pant, P. Srinivasan, and M. E. Ruiz. Evaluating topic-driven web crawlers. In *SIGIR*, 2001.

[22] R. Nayak, P. Senellart, F. M. Suchanek, and A. Varde. Discovering interesting information with advances in Web technology. *SIGKDD Explorations*, 14(2), 2012.

[23] A. G. Parameswaran, H. Garcia-Molina, H. Park, N. Polyzotis, A. Ramesh, and J. Widom. Crowdscreen: algorithms for filtering data with humans. In *SIGMOD*, 2012.

[24] A. G. Parameswaran, A. D. Sarma, H. Garcia-Molina, N. Polyzotis, and J. Widom. Human-assisted graph search: it's okay to ask questions. *PVLDB*, 4(5), 2011.

[25] M. L. Puteman. *Markov Decision Processes*. Wiley, 2005.

[26] T. Risse, S. Dietze, W. Peters, K. Doka, Y. Stavrakas, and P. Senellart. Exploiting the social and semantic Web for guided Web archiving. In *TPDL*, 2012. Poster.

[27] D. Suciu, D. Olteanu, C. Ré, and C. Koch. *Probabilistic Databases*. Morgan & Claypool Publishers, 2011.

[28] R. S. Sutton and A. G. Barto. *Reinforcement Learning*. MIT Press, 1998.

[29] G. Wiederhold. Mediators in the architecture of future information systems. *IEEE Computer*, 25(3), 1992.

Generated Sun Mar 2 12:04:08 CET 2014