On the Convergence of Structured Search, Information Retrieval and Trust Management in Distributed Systems

Karl Aberer, Philippe Cudré-Mauroux, and Zoran Despotovic

School of Computer and Communication Sciences
EPFL, Lausanne — Switzerland
{karl.aberer, philippe.cudre-mauroux, zoran.despotovic}@epfl.ch

Abstract. The database and information retrieval communities have long been recognized as being irreconcilable. Today, however, we witness a surprising convergence of the techniques used by both communities in decentralized, large-scale environments. The newly emerging field of reputation based trust management, borrowing techniques from both communities, best demonstrates this claim. We argue that incomplete knowledge and increasing autonomy of the participating entities are the driving forces behind this convergence, pushing the adoption of probabilistic techniques typically borrowed from an information retrieval context. We argue that using a common probabilistic framework would be an important step in furthering this convergence and enabling a common treatment and analysis of distributed complex systems. We will provide a first sketch of such a framework and illustrate it with examples from our previous work on information retrieval, structured search and trust assessment.

1 Introduction

The database and information retrieval communities have long been perceived as being irreconcilable. The different ways of how data is represented, interpreted and processed are at the core of the divergence in focus of these communities.

The main problem addressed by the database community can be stated as the efficient management of data represented in some first order logic language and the efficient evaluation of queries specifying information needs unambiguously through logical expressions. Recently this model has been extended in the context of the Semantic Web to deal with distributed, heterogeneous information sources by using shared first order conceptual models (ontologies) and a common Web-based infrastructure.

On the other hand, the information retrieval community focuses on finding models for retrieving documents in response to incompletely or ambiguously specified information needs by exploiting document features and user relevance feedback. Web search engines are the most prominent incarnation of these techniques for assessing relevance of documents in response to user requests for information, using both textual content of documents and user feedback derived from the link structure of the Web.

Attempts to reconcile the two communities reach far back in history. Even a conference series, the International Conference on Information and Knowledge Management (CIKM), is dedicated to this goal. We were interested to see to which extent the interaction among the communities progressed, and analyzed the program of the years 2003 and 2004. The result is not too impressive. Among 120 research papers we could identify 10 that are at the borderline of databases and information retrieval, whereas the others are quite clearly belonging to the fields of classical database, information retrieval or knowledge management. In 2004, two sessions on databases and information retrieval have been organized. The topics addressed by the borderline papers are on storage management for retrieval systems, processing of XML documents and similarity search in databases. The last two areas in fact indicate one reason why the boundary between the database view of structured data processing and the information retrieval view of content-oriented processing is starting to dissolve. It is the result of processing specific data types that require both structural and content-oriented processing.

In this paper, we argue that recent developments in diverse areas, such as the Semantic Web, peer-to-peer computing, sensor networks, agent technologies and Web retrieval, indicate that the "semantic gap" between traditional logic-based knowledge presentation and processing and the probabilistic approach taken in information retrieval will be rapidly closing, for a very fundamental reason, that goes beyond the requirement of processing specific data types.

In a distributed environment of autonomous information sources, information and information needs can no longer be expressed concisely, as expected by database and semantic web technologies, but have to deal with *numerous sources of uncertainty*, thus requiring a probabilistic view in the processing of data. In information retrieval, one deals with one specific kind of uncertainty, uncertainty about users information needs. We claim that in distributed environments, qualitatively different sources of uncertainty have to be dealt with as well. This will require a structured framework to represent and process the different sources of uncertainty to provide insightful answers to users information needs. This requirement goes well beyond existing capabilities of both database and information retrieval techniques and systems.

We will illustrate this convergence process by providing several important examples of how the uncertainty resulting from autonomy and incomplete knowledge in distributed environments affects information processing. These examples are taken both from our own work and from some typical results found in the literature. We will provide short summaries of these techniques and illustrate by a simple example of a search problem how each of these techniques affects the information processing task for satisfying the search task. By doing this we illustrate how using a probabilistic framework makes it possible to integrate different ways of dealing with uncertainty, just as first order logic is being used as an integration framework for structured representation and reasoning over distributed information sources. This example-based analysis will allow us to derive some basic conclusions on requirements and issues for extending the current Web infrastructure for dealing with uncertainty in a systematic and integrated way.

2 Running example: Getting newspaper articles about hot days in Switzerland

To illustrate our claims, we introduce an example which is in our opinion representative of the current challenges emerging in information management today. The example starts as a simple SQL query posed against a relational database but will be enriched throughout the paper as new sources of uncertainty are introduced.

From June to August 2003, unusually high temperatures were reported across Europe, including Switzerland. Imagine a journalist wanting to retrieve all newspaper articles about hot days in Switzerland which appeared exactly on one of those days. In a standard relational databases scenario, this could translate to a SQL query like the following:

```
SELECT article.text
FROM articles, weather WHERE
   article.text like %hot summer days%
   and article.date = weather.date
   and weather.temperature  > 30
```

The query contains three predicates, q_1, q_2 and q_3 representing some condition on the content of articles, their publication date and some temperature record respectively.

From a logical perspective, such a query can be considered as a logical expression q for which we have to find all objects d contained in a database such that the implication $d \rightarrow q$ is true.

Expressing an information need in this form reflects several basic assumptions being made, including the ability of the user to precisely express her information need, the correct interpretation of the schematic information provided by the database and the correctness of the data stored in the database. In practice, as we will demonstrate in the following, none of these assumptions can be taken for granted in realistic, distributed information systems.

3 Uncertainty on users' information needs

Since long it has been recognized that logics is not an appropriate framework for information search when it comes to searching documents with textual content. Boolean retrieval has been an early attempt to apply logics for text search, which has soon found its limitations. Due to the ambiguity of natural language, there exists no strict relationship between queries expressed in natural language against documents containing natural language text. Thus the discipline of information retrieval has developed a rich set of models for assessing the relevance of documents for a given query. These models introduce an element of *uncertainty* into the search process, since result objects are no more included into the result set by virtue of a decidable property (a predicate) but whenever there is indication that they might be relevant to some degree to the users information need. These observations clearly apply to the clause q_1 `== article.text like %hot summer days%` of our example query, which in a current database system (ideally) would not be resolved at the syntactic level searching for the exact phrase, but using an underlying text retrieval system.

3.1 Running Example: Accounting for the uncertainty on information needs through probabilistic retrieval

Since we are aiming at a probabilistic framework for dealing with uncertainty in modern information systems, we provide here a short overview of information retrieval from a probabilistic perspective, which follows the exposition given by [5]. From a logical perspective, answering a query q with document d amounts to proving that the implication $d \rightarrow q$ is true. In Boolean retrieval this means that all terms of a (conjunctive) query q would appear in d. In contrast, probabilistic retrieval adopts the following notion for answering a query q: the conditional probability $P(q|d)$ indicates of how relevant document d is to query q.

For computing this probability usually a concept space C of disjoint concepts $c \in C$ is introduced with a probability density function $P(.)$ over C. Queries and documents are considered as concept sets. Then the query answer can be represented as follows:

$$P(q|d) = \frac{P(q \cap d)}{P(d)},\ P(d) = \sum_{c \in d} P(c),\ P(q \cap d) = \sum_{c \in q, c \in d} P(c)$$

A popular type of concepts are terms taken from a vocabulary. Since the concepts are considered as being independent we can further derive

$$P(q|d) = \frac{P(q \cap d)}{P(d)} = \frac{\sum_{c \in C} P(d \cap q \cap c)}{P(d)} = \frac{\sum_{c \in C} P(d \cap q|c)P(c)}{P(d)}$$

If the concept space consists of the terms of a vocabulary, we may assume that the probabilities $P(d|c)$ and $P(q|c)$ are known from analyzing the text collection. For computing a query answer, a standard assumption that is made in probabilistic retrieval is the *maximum entropy principle*, which states the following independence:

$$P(d \cap q|c) = P(d|c)P(q|c).$$

Using this assumption we get

$$\begin{aligned} P(q|d) &= \frac{\sum_{c \in C} P(d \cap q|c)P(c)}{P(d)} \\ &= \frac{\sum_{c \in C} P(d|c)P(q|c)P(c)}{P(d)} \\ &= \sum_{c \in C} P(q|c)P(c|d) \end{aligned}$$

The last expression can be interpreted as the classical model of vector space retrieval, the predominant model for modern text retrieval. Under this interpretation, $P(c|d)$ corresponds to the term weight for a document representation, which is typically computed using a (heuristic) tf-idf scheme and gives the probability that a term is characteristic for a given document. $P(q|c)$ corresponds to the query term weight and gives the probability that a term is characteristic for the result set of query q.

In summary, a predicate such as q_1 `== article.text like %hot summer days%` corresponds in a search model that is considering uncertainty on users' information needs to a random variable q_1 for which we have a method to compute $P(q_1|d)$, the probability that a document is relevant to the predicate. The method to compute this probability relies on an intermediary concept (or feature) space C, for which we assume to have probabilistic models for $P(q_1|c)$ and $P(c|d)$ for a random variable c over the concept space. The computation of $P(q_1|d)$ is then performed by marginalization of the joint probability distribution $P(q_1, c, d)$ exploiting the separation of the random variables q_1 and d through c.

From a practical perspective, using a retrieval engine within a logics-based query language such as SQL poses the question of how to reflect the probabilistic evaluation of $q1$ into the query result. Two solutions are applicable: either only result documents are included that exceed a certain threshold probability. This seems to be problematic with respect to the interpretation of the result. Alternatively the probability values are included into the result table. This might raise efficiency concerns as the result set might become unacceptably large. As we will show in the following, this is a problem that is not confined to the case of dealing with users' uncertainty on information need, but with dealing with uncertainty in general.

4 Uncertainty on knowledge conceptualizations

Traditionally, knowledge representations have been based on subsets of first-order logic in computer science. Indeed, it is widely recognized that knowledge can be efficiently captured by characterizing classes of objects and their inter-relationships. Databases have long used dialects derived from first-order logic to represent or query data, while description logic, a subset of first-order logic, has been chosen to back-up standards for the Semantic Web.

These representations have proven to be extremely useful for dealing with knowledge bases or providing sound semantics to query processing. Until recently, most information-processing tasks took place in controlled environments where one had full control over the definitions of entities in the universe of discourse. When semantic heterogeneity occurred, for examples when multiple schemas or ontologies had to be merged together, some higher-order element (e.g., an integrated schema) was statically introduced to consolidate knowledge in a consistent manner. Thus, some well-known techniques such as Global-As-View and Local-as-View to integrate heterogeneous databases and rewrite queries in deterministic ways have been developed.

Today, however, with the advent of the Internet and the democratization of Semantic Web tools facilitating knowledge elicitation in machine-processable formats, the situation is quickly evolving. One cannot rely on global, centralized schemas anymore as knowledge creation and consumption are getting more and more dynamic and decentralized. In such settings, one has to account for the fact that new knowledge and knowledge representations can appear on a continual basis without any central coordination, while well-known sources might well disappear without prior notice. As a corollary, it is getting more and more difficult to get any kind of certainty about knowledge coming from heterogeneous and dynamic sources over which one has little control.

In this context, uncertainty over knowledge gets particularly critical when one considers agreement on knowledge conceptualizations. Traditionally, only relevant information adhering to specific schemas, taxonomies or ontologies was returned as result of a structured search. As more and more conceptualizations get available from heterogeneous sources, one has to take into consideration the tradeoff between maximizing the precision of the results (by focusing on well-known information sources only) and the total number of relevant results (by considering as many information sources as possible). Many different (semi-) automatic schema mapping schemes have been explored recently. In most cases, some probabilistic value can be returned indicating whether or not the outcome of the mapping process makes sense. One could hence take advantage of these probabilistic values upon deciding whether or not to include an information source for a given structured query.

4.1 Running example: Accounting for the uncertainty on shared conceptualizations through Semantic Gossiping

To come back to our running example, let us imagine that the journalist has access to various newspaper databases on the web. Each database was developed independently of the other ones. All databases consider some sort of representation to encode the date on which a particular newspaper article was published. However, some call this date *published_date*, while other might call it *dateDePublication* or *pd_field*. Due to the fact that the schemas are continually evolving, appearing or disappearing without any central coordination, maintaining a global schema from / to which all individual databases could be mapped is arguably impracticable. Instead, translation links (e.g., schema mappings, views) are defined between pairs of schemas. Those pairwise links permit to iteratively propagate a query posed against a specific schema to other databases. This approach has been taken in the new field of peer-to-peer data management.

The problem lies here in the fact that those links might be created (semi-) automatically, or might not be able to guarantee the outcome of a query mapping deterministically. Different cases may occur in practice. For example, *publication_date* might be erroneously mapped onto *deletion_date* or could be imperfectly mapped onto a *publicationWeek* attribute of a weekly newspaper (coarser degree of granularity for storing publication dates). Thus, we cannot expect the outcome of a query mapping to be one hundred percent faithful to the original query.

We engineered heuristics to quantify the degree to which a translated query differs from the intended query. We termed these techniques *Semantic Gossiping* [1, 2] as they rely on gossiping a query through the various translation links for deriving probabilistic guarantees on the translation process. From a high-level perspective, our methods work as follows: after propagating queries throughout the network of translations, we collect feedback information f, both from the analysis of transitive closures of the query translation processes and from the results received from other databases.

We illustrate how such an approach introduces uncertainty into query answering for one specific type of approach when analyzing feedback received from issuing queries to a peer-to-peer schema mapping network. Given a cycle of mappings $m, m_1, \ldots, m_n$ and assuming all mappings are correct the composite mapping results in a partial identity function. We call this positive feedback f^+. We denote with m_i a random (Bernoulli) variable for a mapping m_i being correct and assume a prior probability ϵ of a mapping m_i being incorrect $P(m_i = 1) = 1 - \epsilon$. Furthermore we assume the probability δ of a mapping error to be compensated in the last step of the cycle by another mapping error to be known. Then we can derive the probability of receiving positive feedback, e.g.,

$$P(f^+|m = 1) = (1 - \epsilon)^n + (1 - (1 - \epsilon)^{n-1})\delta$$

Similarly, other probabilities, e.g. $P(f^+|m = 0, \epsilon, \delta)$ can be computed. We assume that we obtain a set of positive feedbacks $\mathcal{F}^+ = \{f_1^+, ..., f_n^+\}$ and of negative feedbacks $\mathcal{F}^- = \{f_1^-, ..., f_m^-\}$, $\mathcal{F} = \mathcal{F}^- \cup \mathcal{F}^+$ and want to determine the probability $P(m|\mathcal{F})$ of mapping m being correct under these observations. Assuming independence of feedbacks (which in fact is an oversimplification for a real mapping graph) we have

$$P(m|\mathcal{F}) = \prod_{f \in \mathcal{F}} P(m|f).$$

From there, and from the assumption that we have no prior knowledge on m (applying the maximum entropy principle implies $P(m = 1) = P(m = 0)$) we get

$$P(m|f) = \frac{P(f|m)P(m)}{\sum_{m \in \{0,1\}} P(f|m)}$$

Thus, we can determine the conditional probability $P(m|\mathcal{F})$ of a mapping m being correct given some feedback information $\mathcal{F}$. Applying this to our problem, we can determine the probability $P(q_2|\mathcal{F})$ of the date predicate being semantically preserved after applying a mapping m for obtaining the date value, based on feedback information about that mapping:

$$P(q_2|\mathcal{F}) = \frac{\sum_{m \in \{0,1\}} P(q_2, \mathcal{F}|m)P(m)}{P(\mathcal{F})} = \sum_{m \in \{0,1\}} P(q_2|m)P(m|\mathcal{F})$$

making use of the independence assumption $P(q_2, \mathcal{F}|m) = P(q_2|m)P(\mathcal{F}|m)$.

5 Uncertainty on assertions

The quality or pertinence of assertions may greatly vary in decentralized settings. Putting aside trust-related issues (see below for a discussion on this topic), we can expect an ever increasing proportion of automatically-generated assertions in large-scale environments. Fuzzy logic, probabilistic or machine-learning approaches will certainly all contribute at deriving new assertions from existing ones.

Also of interest, the emerging field of sensor networks providing streams of raw data from sensor measurements. Sensors cannot deliver continuous data on extended periods of time due to energy constraints: In fact, there is a well-know trade-off between the precision of sensor data on the one hand, and the battery life of the sensors on the other hand. This implies the necessity of accounting for uncertainty while processing assertions derived from a data acquisition network. The question is, again, how to capture the degree of uncertainty related to the new assertions and how to take advantage of these degrees to get meaningful answers to queries.

5.1 Running Example: Accounting for the uncertainty on sensor measurements in data acquisition networks

Recently, a few probabilistic approaches appeared for processing queries in sensor networks. BBQ [3], for example, introduces the concept of model-based querying. The approach is based on a probabilistic model that captures the correlations among measurements of spatially and temporally correlated sensors, e.g., temperature sensors, to support query answering. The probabilistic model is derived from historical sensor measurements. For query answering, available sensor readings are used to answer user queries by computing the posterior probabilities of the measurement variables from the probabilistic model of the sensor network. In this way missing or faulty readings can be interpolated by the probabilistic model and opportunities for optimizing the physical cost of operating sensor networks can be taken advantage of, such as optimization of energy consumption and reduction of deployment and maintenance cost. We provide in the following a somewhat simplified high-level description of this approach.

Let us assume that the temperatures in `weather.temperature` (q_3) are gathered by a data acquisition network consisting of n fixed sensors, scattered all around Switzerland. They periodically transfer some temperature measurements s_i to a central server. From historical measurements a probability density function $P(s_1, \ldots, s_n)$ is derived. This function captures correlations of temperature measurements due to spatial vicinity of sensors. The model has been extended to also consider temporal correlations. In the case of BBQ this probability density function is a multivariate Gaussian function. The temperature in Switzerland is then defined as the average value of the currently measured values, i.e. $t = \frac{1}{n}\sum_{i=1}^{n} s_i$. If $P(s_1, \ldots, s_n)$ is a multivariate Gaussian, $P(t)$ follows a Gaussian distribution also.

Assume now that a probably incomplete set of raw observations from a subset of all sensors is available, $\mathcal{S} = \{s_j = s_j^o, j \in O\}, O \subseteq \{1, \ldots, n\}$. Then the average temperature can be determined by marginalization as follows

$$P(t|\mathcal{S}) = \int P(s_1, \ldots, s_n|\mathcal{S}) I_t\bigl(\frac{1}{n}\sum_{i=1}^{n} s_i\bigr) ds_1 \ldots ds_n$$

where $I_t(.)$ is the indicator function and

$$P(s_1, \ldots, s_n|\mathcal{S}) = \frac{P(\overline{s}_1, \ldots, \overline{s}_n)}{P(\mathcal{S})}$$

where $\overline{s}_j = s_j^o$ for $j \in O$ and $\overline{s}_j = s_j$ otherwise.

For evaluating predicate q_3 we can derive from

$$P(q_3|t) = \begin{cases} 0 \text{ if } t \leq 30 \\ 1 \text{ if } t > 30 \end{cases}$$

in a now familiar way a probabilistic value for the predicate q_3 being correctly evaluated giving a set of raw measurements s gathered by sensors:

$$P(q_3|\mathcal{S}) = \int \frac{P(q_3, \mathcal{S}|t)P(t)}{P(\mathcal{S})} dt = \int P(q_3|t)P(t|\mathcal{S})dt.$$

6 Reputation-based trust management in decentralized settings

Up to this point we have considered the uncertainties resulting from interpreting factual data (stored in some database) with respect to the intended semantics of a user query. These models exploited intrinsic properties of the data objects being searched for and their associated schemas. These intrinsic properties directly pertain to the query and data objects under consideration. In different applications it can be observed that in addition to these intrinsic features also extrinsic features derived from the context in which the data objects are being used may have an important impact on the search. Trust is a typical example of such extrinsic features. Going back to our running example, we might wonder whether a given article with the content describing hot summer days can be trusted or not. More precisely, only if the newspaper that published the article can be trusted with a sufficiently high probability then we would like to see the article included in the result set.

6.1 Running Example: Accounting for the uncertainty on trustworthiness of the information providers

Imagine the following scenario. An article from a specific newspaper has been reported as containing information on “hot summer days”, so the predicate `q1 == article.text like %hot summer days%` seems to be satisfied. It happened that the user read many articles from that newspaper and was always satisfied with the accuracy of their content. It is intuitively clear that the content from a new article will be accepted by the user. Similarly, the user may use her predominantly negative experiences with the newspaper to conclude that the returned article has to be rejected. Both of these two cases are very extreme in the sense that user *knows* whether to rely on the article content or not; there is little uncertainty here. But the reality is normally somewhere in between.

First, the user may have some positive and some negative experiences with the concerned newspaper. It becomes now unclear whether the predicate is satisfied or not. Second, the user may have never heard about the newspaper, in which case the problem becomes even more severe.

Along the previous discussion, we believe that the problem can be viewed in the following way. Newspapers might be inclined to write in specific ways. For example some may accurately transfer the factual information they collect. Some may exaggerate so that a warm day becomes “very hot.” Some may lie deliberately. The readers can behave similarly when reporting on how they view specific newspapers. Their experiences with the newspaper constitute what we call the newspaper’s reputation. But any given newspaper has many readers and the notion of reputation normally extends to the entire readers community. The readers can share their opinions, even newspapers can write in favor of some other ones etc. So, technically, a whole graph may emerge that encodes the readers’ opinions about newspapers, eventual newspapers’ statements about other newspapers, even readers’ opinions about other readers are possible, they may say a lot about whether a specific reader is bad-mouthing a newspaper for a reason different than the quality of its articles.

There are many approaches that operate on such structures and try to establish trust of the involved entities. In our example this would mean that they can predict how exactly a given newspaper writes. Three fields, web search, semantic web and P2P systems offer good examples of such approaches. [6] presents a well-known technique to rank web pages based on the web link structure. A page is highly ranked if it has many incoming links and/or if the referring pages are themselves highly ranked. The notion of trust is just implicitly present here, in the relative order of the pages. Thus it is hard to talk about a probability of being trustworthy given a link structure. The same holds for [8], which provide a characterization of a class of algorithms to efficiently compute the relative order of the involved semantic statements. In our previous work [4], we establish the link between reputation and trust in the probabilistic sense. We assume that specific joint probability distributions determine the behavior of all involved entities, in our example readers and newspapers, and derive their associated trust as probability distributions over their possible performances.

As a simple example, let us assume that readers report the trustworthiness of the newspapers they happen to read. Thus any newspaper gets associated with a set $\mathcal{R} = \{r_1, r_2, \ldots, r_n\}$, $r_i \in \{0, 1\}$, with the following meaning: ith ($1 \leq i \leq n$) reader claims that the newspaper's trustworthiness is r_i, where 1 stands for "trustworthy" and 0 "untrustworthy." Consider now a reader who wants to make use of this information to decide whether the newspaper can be trusted or not. Having read a number of other newspapers and being able to compare her own opinions about them with those of other readers our reader can assess the probability that the rest of the reader population actually misreports. Let λ denote this quantity. Denoting by θ the unknown probability that the newspaper is trustworthy we can write the probability of receiving the reports $\mathcal{R}$:

$$L(\theta) = [\lambda\theta + (1-\lambda)(1-\theta)]^{\sum_{i=1}^{n} r_i}[\lambda(1-\theta) + \theta(1-\lambda)]^{n-\sum_{i=1}^{n} r_i}.$$

It is also called the likelihood of the sample set $\mathcal{R}$. Note that it is a function of the unknown probability θ only, all other variables are known. We wonder now what θ maximizes $L(\theta)$ given our sample set. This value, denote it θ^*, is called the maximum likelihood estimate of the unknown probability θ. In this example we assumed that the newspapers can be either trustworthy or not. Refinements that cover more outcomes are also possible.

Therefore, trust for a specific newspaper becomes a random Bernoulli variable, denoted by tr and taking values 0 and 1, derived from directly observable reputation reports $\mathcal{R}$. From the maximum likelihood estimation we have a probabilistic model for $P(tr|\mathcal{R})$. Assuming that only results from trusted resources should be included into the result we can state $P(q_1|d, tr) = P(q1|d)$ if $tr = 1$ and $P(q_1|d, tr) = 0$ otherwise. Thus we get making the usual independence assumption $P(q_1, \mathcal{R}|tr) = P(q_1|tr)P(\mathcal{R}|tr)$

$$P(q_1|d, \mathcal{R}) = \frac{\sum_{tr\in\{0,1\}} P(q_1, \mathcal{R}|d, tr)P(tr)}{P(\mathcal{R})} = P(q_1|d)P(tr = 1|\mathcal{R}).$$

7 Search under uncertainty

As illustrated in the previous sections, the example search problem, formulated in a logical framework originally, has a good likelihood to turn into a probabilistic formulation in a distributed setting due to various sources of uncertainty involved in the interpretation of data and user query formulations. Thus, answering the original query, which we formulated as the conjunction of three predicates q_1, q_2, and q_3, results on computing the marginals of a joint probability distribution $P(q, q_1, q_2, q_3, d, c, \mathcal{R}, tr, \mathcal{F}, m, \mathcal{S}, t)$. Finding an answer to the search problem then corresponds to assessing the relevance of the query q when $d, \mathcal{R}, \mathcal{F}, \mathcal{S}$ have been observed. By making independence assumptions on the sources of uncertainty, we can write the joint probability distribution as

$$\begin{aligned}
&P(q, q_1, q_2, q_3, d, c, \mathcal{R}, tr, \mathcal{F}, m, \mathcal{S}, t) = \\
&\quad P(q|q_1, q_2, q_3)P(q_1|c, tr)P(c|d)P(d)P(tr|\mathcal{R})P(\mathcal{R}) \\
&\quad P(q_2|m)P(m|\mathcal{F})P(\mathcal{F})P(q_3|t)P(t|\mathcal{S})P(\mathcal{S})
\end{aligned}$$

The situation can be summarized in a graphical form, e.g., with the Bayesian Network from Fig. 1 below[1]. For each source of uncertainty, we derive a model from a set of observations. The model is in turn used to derive probabilistic guaranties on the predicates of the query being satisfied or not. In the end, the probability on the query being correctly evaluated for a given document and sets of observations $P(q = true|d, \mathcal{R}, \mathcal{F}, \mathcal{S})$ can be computed as

$$\begin{aligned}
&P(q = true|d, \mathcal{R}, \mathcal{F}, \mathcal{S}) \\
&\quad = \sum_{Q_1, Q_2, Q_3} P(q = true|q_1, q_2, q_3, d, \mathcal{R}, \mathcal{F}, \mathcal{S})P(q_1, q_2, q_3|d, \mathcal{R}, \mathcal{F}, \mathcal{S}) \\
&\quad = \sum_{Q_1, Q_2, Q_3} P(q = true|q_1, q_2, q_3)P(q_1|d, \mathcal{R})P(q_2|\mathcal{F})P(q_3|\mathcal{S}) \\
&\quad = P(q_1 = true|d, \mathcal{R})P(q_2 = true|\mathcal{F})P(q_3 = true|\mathcal{S})
\end{aligned}$$

with $P(q_1|d, \mathcal{R})$, $P(q_2|\mathcal{F})$ and $P(q_3|\mathcal{S})$ derived as above, Q_1, Q_2, Q_3 ranging over $\{true, false\}$ for q_1, q_2, q_3 and $P(q = true|q_1, q_2, q_3) = 1$ if $(q_1 = true) \wedge (q_2 = true) \wedge (q_3 = true)$ and 0 otherwise. These derivations can be efficiently handled using well-known techniques such as Belief Propagation or Message-Passing schemes.

Some of our independence assumptions might however not hold in general: for example, trusting (tr) a source might well influence our model on the correctness on its mappings (m) or vice-versa. Also, detection of correct mappings might depend on sensor data, while considering a specific document might be dependant on the trustworthiness of its source, etc. Handling complex conditional relationships between various sources of uncertainty and their models is way beyond the scope of this paper, but might play a crucial role in deriving sufficiently precise heuristics in practice.

[1] Note that various Bayesian Networks can be derived from the aforementioned independence assumptions. For a discussion on causality, we refer the interested readers to [7]

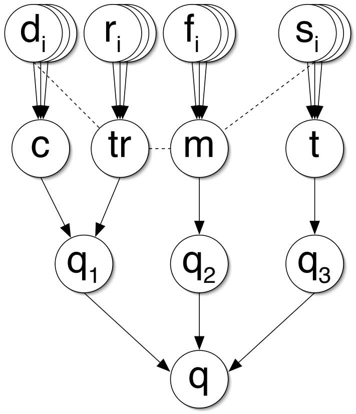

Fig. 1. A Bayesian Network summarizing the conditional dependencies for our running example

8 Conclusions

By now it should have become evident that a systematic treatment of uncertainty in the management of distributed, autonomous information sources will become (or already is) a necessity. We see this as a particularly urgent problem for the emerging field of the Semantic Web which aims at supporting semantically rich information representation for allowing more meaningful information processing, both by humans and machines. Interpretation of data is inherently affected with uncertainty.

A first and critical step for enabling management of uncertainty is the development of and agreement on shared abstractions for representing and handling uncertainty. This is similar to the step that has been taken by the Semantic Web community in agreeing on common logical foundations. Description logics with its many variants has been identified as the proper framework for at least the following reasons. On the one hand it captures the essential elements of conceptual data models used in data management and knowledge representation, on the other hand it provides a computationally tractable framework for reasoning.

Similar issues will have to be taken into account in the search for a common abstraction framework for reasoning under uncertainty. It is a well known fact that complete, probabilistic reasoning is as computationally intractable as reasoning in full first order logic is. AI has a long tradition in developing formalism for reasoning under uncertainty, for example with research lines along Bayesian networks or fuzzy logic. Choosing the proper one has to account for issues of computational feasibility as well as for the possibility to bridge the gap between existing approaches for information processing, such as logical reasoning, machine-learning or information retrieval. We foresee in particular the general extension of usual model-theoretic constructs to take into account uncertainty as an important step to improve structured search results in decentralized settings. This has deep consequences, down to Tarski's Truth definition. The question is: can we provide precise semantics to various probabilistic interpretations in decentralized settings while still developing pertinent, down-to-hearth heuristics for combining or deriving data?

Having selected a proper framework of abstraction, a syntactic representation compatible with existing and evolving Semantic Web standards, such as RDF and OWL, has to be found. This appears to be a comparably trivial task at the first glance. However, a challenge might also be hidden here. As we pointed out earlier, current reasoning techniques for handling uncertainty have typically be developed for isolated problems, and probabilistic statements are consolidated only at the very end of processing queries, as illustrated for our example. As soon as correlations among different aspects of uncertainty are considered, quite surprising problems might occur, which appear to be similar in nature to problems that have been addressed in developing Semantic Web languages, such as RDF, and their processing. How can information on correlations of probabilistic variables, respectively probabilistic statements, be represented in a distributed framework? We can view correlations as the equivalent of relationships, whereas probabilistic variables can be considered as the equivalent of entities. In a distributed setting, managing relationships introduces problems of addressing, assigning responsibilities for storage and management and interoperating with existing infrastructures, all of which would also have to be addressed if probabilistic correlations are managed in a distributed setting.

In summary, we believe that we are seeing today only the very first steps towards an information processing infrastructure that truly accounts for the inherent uncertainty in distributed information processing. Substantial research and development will be required, and both challenging theoretical questions and practical problems have to be mastered. The convergence of developments in different fields such as information retrieval, databases and the Semantic Web will be the main drivers for this development. The reward will be better qualified responses to our ever increasing information needs.

References

1. K. Aberer, P. Cudré-Mauroux, and M. Hauswirth. Start making sense: The Chatty Web approach for global semantic agreements. *Journal of Web Semantics*, 1(1), 2003.
2. K. Aberer, P. Cudré-Mauroux, and M. Hauswirth. The Chatty Web: Emergent Semantics Through Gossiping. In *International World Wide Web Conference (WWW)*, 2003.
3. A. Deshpande, C. Guestrin, S. Madden, J. M. Hellerstein, and W. Hong. Model-Driven Data Acquisition in Sensor Networks. In *Very Large DataBases (VLDB)*, pages 588–599, 2004.
4. Z. Despotovic and K. Aberer. A Probabilistic Approach to Predict Peers' Performance in P2P Networks. In *Eighth International Workshop on Cooperative Information Agents, CIA 2004*, Erfurt, Germany, 2004.

5. N. Fuhr. Models in Information Retrieval. In *European Summer School in Information Retrieval (ESSIR)*, 2000.
6. L. Page, S. Brin, R. Motwani, and T. Winograd. The PageRank Citation Ranking: Bringing Order to the Web. Technical report, Stanford University, Stanford, CA, 1998.
7. J. Pearl. *Causality: Models, Reasoning, and Inference.* Cambridge University Press, 2000.
8. M. Richardson, R. Agrawal, and P. Domingos. Trust management for the semantic web. In *Proceedings of the Second International Semantic Web Conference*, pages 351–368, Sanibel Island, FL, 2003.