Personalized Search for the Social Semantic Web

Oana Tifrea-Marciuska
supervised by Thomas Lukasiewicz, thomas.lukasiewicz@cs.ox.ac.uk
Dept. of Computer Science, University of Oxford
Wolfson Building, Parks Road, Oxford, OX1 3QD, UK
oana.tifrea@cs.ox.ac.uk

ABSTRACT

The Web has recently been changing more and more to what is called the Social Semantic Web. As a consequence, the ranking of search results no longer depends solely on the structure of the interconnections among Web pages. In my research, I argue that such rankings can be based on user preferences from the Social Web and on ontological background knowledge from the Semantic Web, therefore I combine preference representation languages with Semantic Web technologies. Research in database community had dedicated some time to integrate preferences in database queries. In my thesis, as a first step towards closing the gap between the Semantic Web, databases, and preferences, we introduce families of expressive extensions of Datalog$^{\pm}$ with preferences as new paradigms for query answering over ontologies.We first define the syntax and semantic of the proposed frameworks, then propose a top-k query answering algorithm under user preferences in semantic data for different types of queries and preferences models. Each of our proposed frameworks comes with advantages and disadvantages therefore, we provide formal properties of our algorithms and empirical experiments on the performance and quality of our results. Furthermore, we explore the combination of our framework with uncertainty and the generalization to the preferences of a group of users, where we analyze properties of our algorithms related with social choice theory.

1. INTRODUCTION

During the recent years, we are witnessing a change of the Web, from linked Web pages to more (i) semantic data and tags constrained by ontologies, and (ii) social data, such as connections, interactions, reviews, and tags. In the new era of Web, users play an increasingly central role in the creation and delivery of content.

The combination of these two technological waves is called the *Social Semantic Web* (or also *Web 3.0)*. This requires new technologies for search and query answering. The ranking of search results is not solely based on the link structure between Web pages anymore, but on the information available in the Social Semantic Web - in particular, the underlying ontological knowledge present in user–created content, as well as the preferences that the user implicitly or explicitly presents in such content.

The use of semantic search in the Social Web is of central importance, due to the missing link structure between Web pages, which is well-known from ranking (such as PageRank) in standard Web search. In addition, the fundamentally human component of these systems makes each user's personal preferences have a much more prevalent role than what was observed before this paradigm shift. The semantic data can provide precise and rich results, while preferences can help us order the answers of a query.

Finally, the presence of uncertainty in the Web in general is undeniable [4][14]: information integration (as in a travel site that queries multiple sources to find hotels and flights), automatic processing of Web data (analyzing an HTML document often involves uncertainty), as well as inherently uncertain data (such as user comments) are all examples of uncertainty that must be dealt with in answering queries in the Social Web.

The current challenge for Web search is therefore inherently linked to: (1) leveraging the social components of Web content towards the development of some form of semantic search and query answering on the Web as a whole, and (2) dealing with the presence of uncertainty in a principled way throughout the process.

My thesis deals with combining preferences with Semantic data and uncertainty. I analyze the best combination (natural, expressive, concise, efficient, compact) of Semantic Web languages, preference representation languages, and uncertainty to answer personalized queries for a single user or a group of users. This is a challenging task, since on the one hand there are a number of preference representation languages and frameworks in AI and, on the other hand, there are many query languages in the Semantic Web. Combining them requires understanding the advantages and disadvantages of each of these languages.

The goal of my thesis is threefold:

- First, we aim at bridging a gap between Semantic Web languages and preferences representation languages. To do this, we define the syntax and semantic of the three proposed semantic preference frameworks and how to compute top-k answers under these frameworks.
- Second, we aim to understand the advantages and disadvantages of each of these languages, therefore we define formal properties of our algorithms and perform

Proceedings of the VLDB 2015 PhD Workshop

empirical experiments on the performance and quality of our algorithms.

- Last but not least, we would like to generalize the usage of our preferences framework for a group of users and for the presence of uncertainty. These generalizations come with new challenges such as how to handle handle disagreement between users when a group asks a query.

The rest of this paper is organized as follows. In Section 2, we recall some basics on Datalog$^{\pm}$. Section 3 introduces the work done in the area of single user preference modeling, while Section 4 presents the work in the area of preferences for a group of users. Section 5 presents related work and finally, Section 6 summarizes the main results of this thesis and gives an outlook on what stills needs to be achieved.

2. PRELIMINARIES

Datalog$^{\pm}$ [5] is a rule-based formalism that combines the advantages of logic programming in Datalog with features for expressing ontological knowledge. We chose the Datalog$^{\pm}$ language as the ontology language for the Semantic Web, because it is highly flexible and it generalizes many ontology languages such as *DL-Lite* family of Description Logics. Moreover, implementations are also available [12], many tractable subclasses of Datalog$^{\pm}$ have been found, and ideas from the database community related to preferences can be integrated in Datalog$^{\pm}$.

A *Datalog$^{\pm}$ ontology* $O = (D, \Sigma)$, where $\Sigma = \Sigma_T \cup \Sigma_{NC} \cup \Sigma_{EGC}$, consists of a database D, finite set Σ_T of dependencies that can have existential quantification in rule heads set (e.g., $person(X) \rightarrow \exists Y father(X, Y)$), a finite set Σ_{NC} of negative constraints (e.g., $p(X) \wedge q(X) \rightarrow \bot$) and certain equality-generating dependencies set Σ_{EGC} (e.g., $(flight(f1, Dest_1) \wedge flight(f1, Dest_2)) \rightarrow Dest_1 = Dest_2$).

3. SINGLE USER PREFERENCES

Modeling the preferences of a user on the Web has also increasingly become appealing to many companies since the explosion of popularity of social media. The other surge in interest is in modeling uncertainty in these domains, since uncertainty can arise due to many uncontrollable factors. This section analyses the formalism proposed, the semantic properties relevant for these formalism, the algorithmic and complexity aspects and, the implementation and evaluation of these formalism.

3.1 Qualitative Preferences with Uncertainty

The PP-Datalog$^{\pm}$- framework [22] combines preferences and probabilistic uncertainty in Datalog$^{\pm}$ ontologies for a single user, where the preferences of every user are expressed as strict partial orders (SPO: irreflexive and transitive binary relations).

Assuming that more probable answers are in general more preferable, one asks how to rank answers to a user's queries, since the preference model may be in conflict with the preferences induced by the probabilistic model - the need thus arises for preference combination operators.

Our work [27] in this area has been on defining new preference combination operators, which produce a preference relation, given a general SPO that models preferences and a score-based SPO that models uncertainty. We have proposed four specific algorithms for such an operator, and analyzed their semantic and computational properties. For example, one operator allows the user to choose how much influence the probabilistic model has on the output. Another example, is to use as base the user's preferences and to use the probabilistic model as a secondary source of "advice".

Semantic Properties. We proved that each of the proposed combination operators return an SPO and if there is no disagreement between the probabilistic model and the preference model, then the combination operator should keep the ordering. These are two reasonable properties that one would expect. Other examples of proved properties related with extreme cases and the presence of cycles.

Algorithms and Complexity. We have studied a basic algorithm for answering k-rank disjunctions of atomic queries, and showed that under certain conditions, k-rank queries can be answered in polynomial time in the data complexity, which is the same complexity as answering traditional preference-based queries in relational DBs. Moreover, we have provided upper bounds of the complexity results for each of the operators.

Implementation and Evaluations. In the case of PP-Datalog$^{\pm}$, we have evaluated and analyzed the running time of our algorithms over a combination of real-world and synthetic data. We use the IMDb for the probabilistic model and the semantic data, while the user preferences were randomly generated using a density factor. The code and dataset are available as open source [28].

3.2 CP-Nets

CP-nets [3] are a graphical representation language that provide a natural, concise, and flexible representation of qualitative preferences and of conditional preferences (e.g., "if meal is dinner, then I prefer to drink wine to tea"). The formalism is based on the 'all other things being equal" semantics (e.g., the dessert, if it is the same in both meals, is irrelevant to our preference on the drink). We have introduced ontological CP-nets [6, 7, 8], which are a novel combination of Datalog$^{\pm}$ ontologies with CP-nets. We have defined CP-nets–based conjunctive queries and their skyline and k-rank answers on top of ontological CP-nets. We further extended [9] the framework with more general preferences, CP-theories[34].

Algorithms and Complexity. We provided an algorithm for computing skyline and k-rank answers to CP-nets–based conjunctive queries. Furthermore, we have provided precise complexity results as well as tractability results for these problems when the underlying ontologies are defined via existential rules.

4. GROUP OF USERS PREFERENCES

Clearly, social media are a valuable source for mining preferences and opinions of groups of users for commercial or political purposes. Therefore, search for a group of users is an important problem.

To address this problem, a model of preferences of individual users can be adopted and then the individual preferences can be aggregated into a group's preferences. However, this comes along with two additional challenges. The first challenge is to define a group preference semantics that solves the (all but certain) *disagreement* among users (a system should return results in such a way that certain properties are satisfied, e.g., ensuring that each individual benefits from the result). E.g., people (even friends) often have different tastes in restaurants. The second challenge is to allow for *efficient algorithms*, e.g., to compute efficiently the answers to queries under aggregated group preferences. Similarly with the previous section we analyze the formalism proposed, the semantic properties relevant for these formalisms, the algorithmic and complexity aspects, and the implementation and evaluation of these formalism.

4.1 Qualitative Preferences

We introduced two ontology languages [24, 25, 26, 23], which combine the Datalog$^{\pm}$ ontology language with group preferences (a generalization of preference handling in relational databases), called GP-Datalog$^{\pm}$, and which combine both preferences of a group of users and probabilistic uncertainty, called GPP-Datalog$^{\pm}$. To our knowledge, these are the first combinations of ontology languages with group preferences with and without probabilistic uncertainty. We provide aggregation operators that take as input group preferences models and compute an aggregated model from all individual preferences.

Semantic Properties. We also present several ways to compute group preferences as an aggregation of sets of single-user preferences, based on social choice theory.

Algorithms and Complexity. We give algorithms for answering k-rank queries for disjunctions of atomic queries, which generalize top-k queries based on the iterative computation of classical skyline answers. We show that answering disjunction of atomic queries in GPP-Datalog$^{\pm}$ and GP-Datalog$^{\pm}$ is possible in polynomial time under certain conditions.

Implementation and Evaluation. The feasibility of the approach is demonstrated by applying it to preference models obtained from real users.

When building a new formalism, one of the problems that arises is the lack of datasets to test empirically your formalism. Since preferences are private information, this data is very difficult to find. Therefore, we gathered the preferences of 50 users, using a web application. Users stated their preferences as SPOs, over cuisine, type of food, and type of place (breakfast, lunch, and dinner). Users entered their preferences (e.g., prefer Italian food over Japanese food). Based on the fact that users know each other, we created 19 groups of 3 to 9 users.

The code of our algorithms and dataset are available as open source [28].

All runs were carried out using the Datalog$^{\pm}$ ontology built from the Yelp Dataset Challenge[35], which contains 11,537 businesses in the Phoenix (USA) metropolitan area. The queries represent situations where groups wish to decide where to go for a meal.

We discussed three research questions: "Q1: *Which approach is more efficient?*" ,"Q2: *Which aggregation method yields the best results?*"', and "Q3: *How different are the results produced by each aggregation method?*".

5. RELATED WORK

Modeling and dealing with preferences in databases has been studied extensively; see [32, 16] for a survey and [17, 19] for a more recent work. Work has also been carried out in the intersection of databases and knowledge representation and reasoning, such as preference logic programs [13], incorporation of preferences into formalisms such as answer set programs [2], answering k-rank queries in ontological languages [21], and combination of Semantic Web technologies with preference representation and reasoning [20, 31].

Many studies address the area of group modeling. Indirectly, it was studied in mathematics or economics[33]. Current approaches that deal with group preferences have also been studied in the area of recommender systems [1, 30], which focus on quantitative preferences.

My thesis aims at making a more formal analysis of the ways we could do personalized search. To best of our knowledge, this is something that is definitely missing in this area: a better understanding on how to combine preferences representation languages, uncertainty, and the Semantic Web.

6. SUMMARY AND OUTLOOK

We analyzed different preferences languages from artificial intelligence and explored which of them are more fit for integration with the Semantic Web. We analyzed both more simple qualitative preferences (e.g., "I prefer wine over pasta") but as well more complex structures (e.g, "If I watch Titanic, then I prefer wine over pasta"). We compared them formally (e.g., complexity of the algorithms, properties) and empirically.

Figure 1 presents the summary of the formalism proposed and the work in progress. Two directions of my work are under development. First, I am working on using score-based preferences on the Social Semantic Web, since the algorithms used in [27] are not suitable for score-based preferences, as they don't leverage this simpler structure. We develop an algorithm for top-k query answering in this framework for a union of conjunction queries with safe negation and to generalize the above approach to top-k query answering under the preferences of a group of users (rather than a single user), which involves the aggregation of (potentially conflicting) user preferences. Second, we intend to provide experimental results on the performance and quality of our algorithms, where it is missing. Another interesting topic for future research is to generalize the presented approach to ontologies with uncertainty.

ACKNOWLEDGMENTS

This work was supported by the EPSRC grant "PrOQAW: Probabilistic Ontological Query Answering on the Web", by the European Research Council grant "DIADEM", by a Google European Doctoral Fellowship, and by a Yahoo! Research Fellowship.

7. REFERENCES

[1] S. Amer-Yahia, S. Roy, A. Chawla, G. Das, and C. Yu. Group recommendation: Semantics and efficiency. *Proc. VLDB Endow.*, 2(1):754–765, 2009.

[2] G. Brewka. Preferences, contexts and answer sets. In *Proc. of ICLP*, p. 22, 2007.

[3] C. Boutilier, R. I. Brafman, C. Domshlak, H. H. Hoos and D. Poole. *CP-nets: A Tool for Representing and Reasoning with Conditional Ceteris Paribus Preference Statements.* J. Artif. Intell. Res. (21) 135-191, 2004

Formalism	Uncertainty	Implementations	Queries	Complexity	Properties	Preferences
PP-Datalog$^{\pm}$ [27]	Yes	Yes [28]	Disjunction of atomic queries	Yes	Yes	Qualitative
GP-Datalog$^{\pm}$ [24, 26]	No	Yes [28]	Disjunction of atomic queries	Yes	Yes	Qualitative
GPP-Datalog$^{\pm}$ [25, 23]	Yes	Yes [28]	Disjunction of atomic queries	Yes	Yes	Qualitative
Ontological CP-nets [6, 7, 8]	No	No	Conjunctive queries	Yes	No	CP-nets
Ontological CP-theories [9]	No	No	Conjunctive queries	Yes	No	CP-theories
Score-Datalog$^{\pm}$	No	In progress	Conjunctive queries	Yes	Yes	Quantitative
Relational Preferences	No	In progress	In progress	In progress	In progress	CP-theories

Figure 1: Summary of my research at the moment

[4] Jung, J. C. and Lutz, C. 2012. Ontology-based access to probabilistic data with OWL QL. In *Proc. of ISWC*. LNCS, vol. 7649. Springer, 182–197.

[5] A. Calì, G. Gottlob, and T. Lukasiewicz. A general Datalog-based framework for tractable query answering over ontologies. *J. Web Sem.*, 14:57–83, 2012.

[6] T. Di Noia,T. Lukasiewicz, M. V. Martinez, G. I. Simari, and O. Tifrea-Marciuska. Ontological CP-Nets. In *Proc. of URSW III*, 2014.

[7] T. Di Noia, T. Lukasiewicz , M. V. Martinez, G. I. Simari, and O. Tifrea-Marciuska. Computing k-Rank Answers with Ontological CP-Nets In *Proc. of SEBD*, 2014.

[8] T. Di Noia, T. Lukasiewicz, M. V. Martinez, G. I. Simari, and O. Tifrea-Marciuska. Computing k-Rank Answers with Ontological CP-Nets In *Proc. of PRUV*, 2014.

[9] T. Di Noia, T. Lukasiewicz, M. V. Martinez, G. I. Simari, and O. Tifrea-Marciuska. Combining Existential Rules with the Power of CP-Theories In *Proc. of IJCAI*, To appear, 2015.

[10] R. Fagin. Combining fuzzy information from multiple systems. *J. Comput. System Sci.*, 58(1):83–99, 1999.

[11] R. Fagin, A. Lotem, and M. Naor. Optimal aggregation algorithms for middleware. *J. Comput. System Sci.*, 66(4):614–656, 2003.

[12] G. Gottlob, G. Orsi and A. Pieris. Query Rewriting and Optimization for Ontological Databases. *ACM Trans. Database Syst.*, 39:1–25, 2014.

[13] K. Govindarajan, B. Jayaraman, and S. Mantha. Preference logic programming. In *Proc. of ICLP*, pp. 731–745, 1995.

[14] Lukasiewicz, T., Martinez, M. V., Orsi, G., and Simari, G. I. . Heuristic ranking in tightly coupled probabilistic description logics. In *Proc. of UAI.* AUAI, 554–563, 2012.

[15] I. F. Ilyas, W. G. Aref, and A. K. Elmagarmid. Supporting top-k join queries in relational databases. *VLDB J.*, 13(3):207–221, 2004.

[16] I. F. Ilyas, G. Beskales, and M. A. Soliman. A survey of top-k query processing techniques in relational database systems. *ACM Comput. Surv.*, 40(4):11, 2008.

[17] M. Jacob, B. Kimelfeld, and J. Stoyanovich *A System for Management and Analysis of Preference Data.* In *Proc. of VLDB*, 2014.

[18] C. Li, K. C.-C. Chang, I. F. Ilyas, and S. Song. RankSQL: Query algebra and optimization for relational top-k queries. In *Proc. of SIGMOD*, 2005.

[19] J. Li, S. Barna, and D. Amol. RankSQL: A unified approach to ranking in probabilistic databases. In *Proc. of VLDB*, 2011.

[20] T. Lukasiewicz and J. Schellhase. Variable-strength conditional preferences for ranking objects in ontologies. *J. Web Sem.*, 5(3):180–194, 2007.

[21] T. Lukasiewicz, M. V. Martinez, and G. I. Simari. Preference-based query answering in Datalog+/– ontologies. In *Proc. of IJCAI*, pp. 501–518, 2013.

[22] T. Lukasiewicz, M. V. Martinez, and G. I. Simari. Preference-based query answering in Datalog$^{\pm}$ ontologies. In *Proc. of IJCAI*, 2013.

[23] T. Lukasiewicz, M. V. Martinez, G. I. Simari, and O. Tifrea-Marciuska. Ontology-based query answering with group preferences. *ACM Trans. Internet Technol.*, 14(4):25, 2014.

[24] T. Lukasiewicz, M. V. Martinez, G. I. Simari, and O. Tifrea-Marciuska. Query Answering in Probabilistic Datalog$^{\pm}$ Ontologies under Group Preferences In *Proc. of WI*, 2013.

[25] T. Lukasiewicz, M. V. Martinez, G. I. Simari, and O. Tifrea-Marciuska. Query Answering in Datalog$^{\pm}$ Ontologies under Group Preferences and Probabilistic In *Proc. of DMSSW*, 2013.

[26] T. Lukasiewicz, M. V. Martinez, G. I. Simari, and O. Tifrea-Marciuska. Group Preferences for Query Answering in Datalog$^{\pm}$ Ontologies In *Proc. of SUM*, 2013.

[27] T. Lukasiewicz, M. V. Martinez, G. I. Simari, and O. Tifrea-Marciuska. Preference-Based Query Answering in Probabilistic Datalog$^{\pm}$Ontologie. Journal on Data Semantics., 1861-2032, 2014.

[28] Personalised semantic search - source code and datasets, 2014. `https://github.com/personalised-semantic-search`

[29] A. Natsev, Y.-C. Chang, J. R. Smith, C.-S. Li, and J. S. Vitter. Supporting incremental join queries on ranked inputs. In *Proc. of VLDB*, 2001.

[30] E. Ntoutsi, K. Stefanidis, K. Nørvåg, and H.-P. Kriegel. Fast group recommendations by applying user clustering. In *Proc. of ER*, 2012.

[31] U. Straccia. Top-k retrieval for ontology mediated access to relational databases. *Inform. Sciences*, 198:1–23, 2012.

[32] U. Straccia. A Survey on Representation, Composition and Application of Preferences in Database Systems. *ACM Trans. Database Syst.*, 19:1–:45, 2011.

[33] A. D. Taylor. *Social Choice and the Mathematics of Manipulation.* Cambridge University Press, 2005.

[34] N. Wilson. *Extending CP-nets with stronger conditional preference statements.* In *Proc. of AAAI*, 2004.

[35] Yelp. Yelp Dataset Challenge, 2012. `http://www.yelp.co.uk/dataset_challenge`