Induction of user preferences for semantic web

Alan Eckhardt

Doctoral thesis, Charles University Prague, 2010
Review of thesis supervisor

Main motivation of the work comes from considering the problem of a user searching the web (and when key word search is not satisfactory). Search should be personalized to a possibly arbitrary user, process should integrate information from several sources, answers have to be ordered according to user preference and user will probably consider only few results on top 1-4 pages. R. Fagin came with a top-k query model and an optimal query answering algorithm for such a situation.

Problem is that Fagin's model does not say how to adapt to a new user, he presents just a model of user preferences consisting of attribute preferences (local user preferences - LUP) and an aggregation (global user preference - GUP), all based on fuzzy (many valued) logic.

So a problem arose, how to find out user preferences. Specific to this approach is a small number of user inputs and a small number of rating (fuzzy) values. This with huge number of data makes preference learning (PL) for top-k querying especially hard.

Semantic web is here implicitly assumed in having objects (resources) identified and attributes annotated. One can imagine, they come from different web locations either as a result of web information extraction or they are annotated e.g. via RDFa, microformats or microdata and transferred through GRDDL and/or a semantic web service. Hence semantic web assumption makes possible to deal with web data as if they were locally integrated (both in our and Fagin's case). Nevertheless, administrative proximity remains large and user web querying has its specifics which can not be overcome even by the semantic web assumption.

Induction of user preferences for semantic web enabling top-k querying is, by my opinion, an up to date, hard and a challenging problem of general interest.

Thesis of Alan Eckhardt under review deals with the above mentioned problem of induction (learning) of user preferences for (semantic web integrated) data top-k querying.

Area of web object search (e.g. product search) is very wide - author restricts to content based search given (single but arbitrary) user rating. This approach is especially suitable for cold start search when there are no data on other users or until user is identified with a group of (similar) users.

Introduction contains motivation and different aspects of problem setting. Well informed related work describes first preference models and then results on preference learning (PL) and user input. Chapter 3 describes in detail the preference model used in this work (suitable for R.Fagin's top-k query evaluation algorithms).

Contributions are described in further chapters: 4 - user preference model learning; 5 – evaluating user preference model learning; 6 - implementation of PrefWork, a framework for PL and testing; 7 – experiment s and an Apendix with details of experiments.

Main contributions fall into several groups – new methods, enhancement of other methods, finding right measures, implementation and experiments.

Author introduced several new methods of PL, especially one called "Statistical" and second "Instance". Even without further testing and comparison, these methods are interesting because they are computationally easy and results are both LUP and GUP (suitable for Fagin's querying). This is especially useful as some traditional methods (e.g. best performing SVM) are computationally demanding. Another result which I value highly is a new method for 2CP regression (Algorithms 1 and 2 on page 48), although it probably needs some more tuning.

Advantage of having LUP is that the data space becomes normalized to $[0,1]^n$ and monotone, with (1, 1, ..., 1) as an ideal point. Many methods (e.g. collaborative filtering CF, UTA ...) were significantly improved by author's new PL methods. LUP understood as data preprocessing for UTA brought improvement. LUP+GUP as a user profile enables similarity extension and improvement in CF. Especially for CF this is a surprise for me.

Experiments have shown that candidate was able to overcome main obstacles of PL – small number of rated object (small training set is a problem of most statistical evaluations) and small size of rating values (witch makes class of most preferred objects very big). Most of tests were run in large number through all possible combination of data x many different users generated x size of rated set (from 5-50) x methods (new and standard) x several measures. This big number of experiments enabled to test statistical significance of results of type "method 1 performed better than method 2" with $p <$ some threshold. Small numbers of rated objects was overcome by bootstrapping and reverse cross validation. Small numbers of rating values was solved by method "Instances" (see Fig.4.3. and related). Results were tested according to several order sensitive measures and different user models. Especially the "natural user model" of Zimmermann-Zysno coming from human user studies is valuable. Only what remains to do, but it was outside the scope of this thesis, are more extensive experiments with real human users (and it needs cooperation with sociology and psychology). Thesis of A. Eckhardt can be seen as a contribution to software engineering as an experimental science. Experiments are all repeatable.

Implementation is nice, can communicate with external methods (e.g. ones from Weka, understood as PL, although often not suitable for Fagin's querying model) and is publicly available. It is a contribution to software testing frameworks.

Concluding, A. Eckhardt has developed (computationally simple) preference learning methods, suitable for underlying indexing and query processing (e.g. Fagin's algorithms), which in many cases significantly outperformed (some with $p < 0.001$) publicly available methods according to several order sensitive measures and are well suited for web search with small number of user ratings and small size of the rating scale.

Alan Eckhardt has proved himself as a reliable and adventurous student fighting in an unknown area in a challenging situation with success and has made a difference.

Thesis presents a high quality improvement of software engineering in the area of preference learning and experimental validation. I recommend thesis for public defense and after successful defense the author to be awarded with Ph.D. title.

Prague, July 24th, 2010

Peter Vojtáš
thesis supervisor