Delivering a Personalized Result Set by the Adaptation of Preference Queries

Sven Döring, Annette Eberle, Timotheus Preisinger

Chair for Databases and Information Systems
University of Augsburg
Universitätsstr. 14
86159 Augsburg
{Doering, Eberle, Preisinger}@informatik.uni-augsburg.de

Abstract: Personalization includes the adaptation of database queries according to the user's needs, wishes and situation. We examine the influence of the d-parameter as powerful personalization instrument for the Preference XPath search engine. Using a heuristic approach we present a possibility to deliver not only the qualitative best matching objects but also the desired amount of data to the user. Performing a series of test queries on proper e-catalog data, we demonstrate the effectiveness of our approach.

1 Introduction

There is an increasing interest in personalization issues of computer applications today [Fo05]. Intuitive interaction between human and computer is absolutely necessary in particular to address people who are inexperienced, handicapped or even afraid of using a computer. Therefore a computer should adapt itself to human needs and preferences so that users do not need to get accustomed to the computer as it is still common today. Any service has to be tailored to the individual customer in order to be successful [Dz04]. Personalization of human-computer interaction comprises various areas of computer science, e.g. artificial intelligence, speech processing, and database technology. Search engines are often the crucial link between user and the system's information, e.g. the product database of e-procurement platforms. Inexperienced users as well as experts prefer search results adjusted to their personal preferences [KI05]. Unfortunately, common search technologies do not respect the individual user's wishes [KFD04].

The size of the search result is as important as the quality of the matches in terms of personalization. For example, a customer of an e-procurement platform may prefer a handy set of 10 articles. Another customer might prefer a set of 6 articles for comparison purposes due to the purchasing policy of his company. Top-k queries are aimed to deliver the desired amount of search results [BGM02]. But using a ranking approach no human comprehensible information about the search result's quality is available. Which user really understands the statement "the result matches your preferences with 70%" delivered by a complex, weighted mathematical computation? Though, an argumentation supported by information about the quality of search results (e.g. "the color is perfect matched") is an important factor, e.g. in convincing a customer.

2 Preference XPath search engine

Taking the user's preferences into account is a promising way for the personalization of database queries [Ch03]. A preference search technology like Preference XPath avoids the annoying empty result effect and reduces the flooding effect with its lots of irrelevant results. The underlying query model delivers best matches only (BMO) wrt. the user's search preferences [Ki02]. Thus, an improved search result quality respecting the individual user's preferences is achieved [Ki05]. Human comprehensible information about the quality of the search results supports the user [KFD04].

In [Ki05] a preference search enabling the adjustment of the search result's size is presented. The so called *d-parameter* for numerical base preferences is introduced as one instrument to personalize the search result's size. The d-parameter allows a partitioning of numerical domains. For example, let us consider a section of a car vendor's database in Table 1. One of his customers might have the wish to get a car with a price around $20.000. This simple wish can be easily transformed in the following Preference XPath query:

/cars #[Price around 20000]#

Please note soft constraints are scoped by '#[…]#'. For the example query above tuple t_2 would be delivered since it is the best alternative. However, some customers might also want to get t_3 and t_5 because they do not care about a price difference of up to $400. Using the d-parameter with a value of 400 the search query could be adequately personalized:

/cars #[Price around (20000, **400**)]#

This time t_2, t_3, and t_5 are delivered because the price domain is partitioned by the d-parameter. Thereby, the prices $20.100, $20.200, and $20.300 would be treated as equal good in terms of this user preference. For detailed information about d-parameter see [Ki05].

	t_1	t_2	t_3	t_4	t_5
Price ($)	23.500	20.100	20.300	22.000	20.200

Table 1: Database of cars

3 Adaptation of preference queries

With the d-parameter the preference search offers a powerful instrument enabling the adaptation of preference search queries for the individual user. In this section we present an empirical examination of the d-parameter. In particular we address the question how to adjust the d-parameter of a numerical base preference in order to deliver a result set of a preferred size. We will show a heuristic approach to handle this question.

3.1 Towards an heuristic query adaptation mechanism

First, we generated electronic product catalogs in XML format by means of the statistic software R [VS02]. We constructed normally and uniformly distributed data. This is sufficient for evaluation purposes since we consider huge amounts of data which can usually be regarded as normal distributed by approximation ([Ka86]). Due to space restrictions in this work we will present our heuristic for normal distributed values only. For example see the normal distribution of the attribute *price* (part A of Figure 1) leading to an expected value of $\mu = 13502.53$ and a deviation of $\sigma = 1512.02$.

A huge number of queries is systematically created and evaluated. We focus on examining the AROUND_d base preference [Ki05]. Using various AROUND_d parameters, the d-parameter is running from 0 to 1.400. We choose this range due to practical reasons. For d=1.400 already up to 1.600 results are delivered – far more than manageable in reality. For each value of the d-parameter the size of the result set is of interest. Obviously, the size of the result set increases with a growing d-parameter (see part B of Figure 1). We approximate the curves of Figure 1 part B by straight lines using a regressions analysis of R [VS02]. Mathematically, these straight lines look like:

$$H_{param}(d) = s_{param} \cdot d + t_{param} \qquad \textbf{[F1]}$$

with a result set's size of $\mathbf{H_{param}(d)}$, a gradient $\mathbf{s_{param}}$, d-parameter **d,** and axis intercept $\mathbf{t_{param}}$**.** This equation can be transformed to calculate the d-parameter to a desired result set size. It is possible to estimate s_{param} and t_{param} doing a computation in advance. Preference queries for different d-parameters and the expected value μ as AROUND_d parameter are executed, counting up d from 0 to δ representing the d-parameter delivering 60% of the data, which is already more than manageable in practical use. We approximate the resulting curve defined by $H_\mu(d)$ with R determining a straight line called $\mathbf{Line_{app}}$. Facts of interest are gradient $\mathbf{s_\mu}$, result set size $\mathbf{H_\mu}$ for d=δ, and $\boldsymbol{\delta}$ itself.

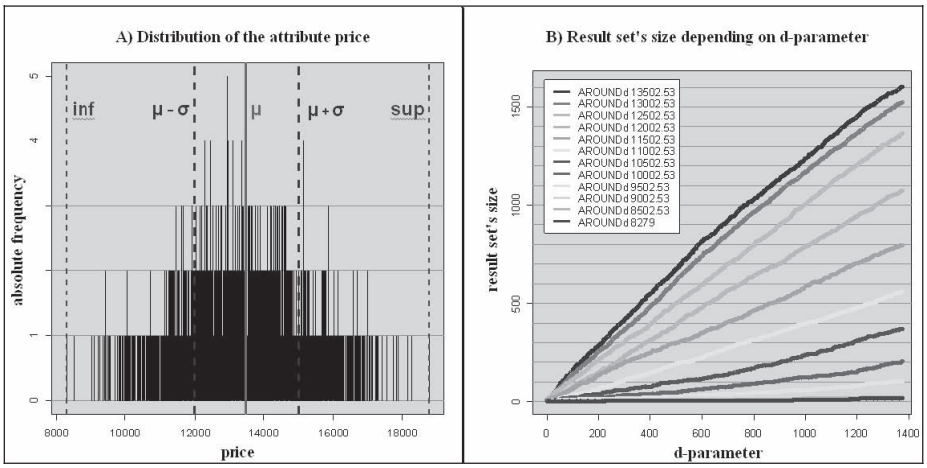

Figure 1: Normal distribution of price (A) and result set size depending on d-parameter (B)

Using the gradient s_μ of $Line_{app}$ it is possible to determine the gradient of approximated straight lines for any $AROUND_d$ parameter. Based on the density function of normal distributions we empirically identify the correlation by

$$s_{param} = s_\mu \cdot e^{-\frac{1}{2\tilde{\sigma}_1^2}(param-\mu)^2} \quad \textbf{[F2]}$$

where: *param* is the $AROUND_d$ parameter; $\tilde{\sigma}_1 = \sigma * (1+25/189)$

This equation of s_{param} and the point $P = (\delta, \hat{H}_{param,\delta})$ consisting of δ as the computed value for the d-parameter (see above) and the result set size $\hat{H}_{param,\delta}$ determined by δ and the according $AROUND_d$ parameter are used in order to get t_{param}. This leads to:

$$t_{param,\delta} = \hat{H}_{param,\delta} - s_{param} \cdot \delta \quad \textbf{[F3]}$$

Again based on the density function of normal distributions we estimate:

$$\hat{H}_{param,\delta} = H_\mu \cdot e^{-\frac{1}{2\tilde{\sigma}_2^2}(param-\mu)^2} \quad \textbf{[F4]}$$

where: $\tilde{\sigma}_2 = \sigma * (1+25/252)$

Please note that the determination of the distribution as well as the computation of δ, σ, μ, s_μ, and H_μ can be done in advance, e.g. while spooling in a new e-catalog into a database. If the user executes a search query, e.g. to find a particular amount of a desired product, the d-parameter of the corresponding preference would be computed on the fly by our following heuristic approach:

given: desired result set size **H,** attribute **price,** and $AROUND_d$ parameter **param**

searched: d-parameter $\mathbf{d_{param,\delta}(H)}$

$$d_{param,\delta}(H) = \frac{1}{s_\mu \cdot e^{-\frac{1}{2\tilde{\sigma}_1^2}(param-\mu)^2}} \left(H - H_\mu \cdot e^{-\frac{1}{2\tilde{\sigma}_2^2}(param-\mu)^2} \right) + \delta \quad \textbf{[F5]}$$

The calculated value for the d-parameter can now be used to adjust a preference query:

\cars #[Price around param, $\boldsymbol{d_{param,\delta}(H)}$]#

Please note that our heuristic approach can also be used for the adjustment of the numerical preferences $LOWEST_d$, $HIGHEST_d$, and $BETWEEN_d$ with only marginal changes due to the preferences' sub-constructor hierarchy of [Ki05].

3.2 Evaluation

We evaluated our heuristic mechanism using a product catalog of SSI-Schäfer (www.ssi-schaefer.com) with around 700 entries per attribute. After determining the distribution of an attribute expected value, deviation, and the approximated straight lines were computed. In the following we started sample queries comparing desired result set size with the result set size delivered by the adjusted search query using the computed d-parameter of our heuristics. Usually, the difference was less than 20%. This difference is caused by approximations. However, this is quite sufficient since a user who prefers a handy result set of 20 is usually satisfied with 16 to 24 results as well.

4 Summary

We presented a heuristic approach for numerical base preferences which is able to adapt the preference search query in order to sufficiently deliver the desired amount of results. Thereby, the user does not only get best fitting results but also the preferred amount achieving a novel level of personalization. This approach can also be used for the adaptation to the user's situation, e.g. to deliver more results to a bigger screen. An adequate modeling of the situational context and the storing of user's preferences are presented in [HK04]. A smart presentation component [KFD04] delivering personalized and situated information about the result's quality represents another beneficial functionality.

Bibliography

[BGM02] Bruno, N.; Gravano, L.; Marian, A.: Evaluating Top-k Queries over Web-Accessible Databases. In Proc. of the Int. Conf. on Data Engineering (ICDE'02), San Jose, USA, 2002; pp. 369-378.

[Ch03] Chomicki, J.: Preference Formulas in Relational Queries. In ACM Transactions on Database Systems (TODS), volume 28, issue 4, 2003.

[Dz04] Dziarstek, C. et. al.: A User-Aware Financial Advisory System. In Proc. of Multikonferenz Wirtschaftsinformatik (MKWI), Berlin, 2004; pp. 217-229.

[Fo05] FORSIP: http://www.forsip.de, requested at 04/01/2005.

[HK04] Holland, S.; Kießling, W.: Situated Preferences and Preference Repositories for Personalized Database Applications. In Proc. of the 23rd Int. Conf. on Conceptual Modeling, Shanghai, China, 2004; pp. 511-523.

[Ka86] Kachigan, S.: Statistical Analysis. Radius Press, New York, 1986.

[KFD04] Kießling, W.; Fischer, S.; Döring, S.: COSIMA B2B – Sales Automation for E-Procurement. In Proc. of the 6th IEEE Conf. on E-Commerce Technology, San Diego, USA, 2004; pp. 59-68.

[Ki02] Kießling, W.: Foundations of Preferences in Database Systems. In Proc. of the 28th Int. Conf. on Very Large Data Bases, Hong Kong, China, 2002; pp. 311-322.

[Ki05] Kießling, W.: Preference Queries with SV-Semantics. In Proc. of the 11th Int. Conf. on Management of Data, Goa, India, 2005; pp. 15-26.

[KI05] Koutrika, G.; Ioannidis, Y.: Personalized Queries under a Generalized Preference Model. In Proc. of the 21st Int. Conf. on Data Engineering, Tokyo, Japan, 2005; pp. 841-852.

[VS02] Venables, W.; Smith, D.: An Introduction to R. Network Theory Ltd, 2002.