University of Zurich[UZH]

Department of Informatics

University of Zürich
Department of Informatics
Binzmühlestr. 14
CH-8050 Zürich
Phone. +41 44 635 43 11
Fax +41 44 635 68 09
www.ifi.uzh.ch/dbtg

UZH, Dept. of Informatics, Binzmühlestr. 14, CH-8050 Zürich

Francesco Luminati

Prof. Dr. Michael Böhlen
Professor
Phone +41 44 635 43 33
Fax +41 44 635 68 09
boehlen@ifi.uzh.ch

Zürich, March 13, 2012

Master Basismodul in Informatik
Datenbanktechnologie

Topic: KNN-Join to compute derive attributes in the Swiss Feed Data Warehouse

The Swiss Feed Database contains chemical parameters of 155 nutrients. This data is available for more than 600 animal feed types and is used by companies private farmers and research institutions to produce healthy, effective and cheap animal feed. Depending on the nutrient, chemical parameters are derived in two ways. Firstly, parameters are measured through chemical analyses on field samples of different animal feeds. Secondly, parameters are computed from measurements of other nutrients.

The computation of derived nutrient parameters is based on known dependencies that are formalized with a help of algebraic expressions, aka, regressions. The complexity of regressions varies depending on the number of involved nutrients. In one case, a regression involves measurements of only one nutrient, in another case, measurements of many nutrients are required to compute the regression. Furthermore, regressions might be defined recursively, i.e., based on the output of other regressions. In all cases, the large number of available regressions makes it hard to manually update the Feed Database as new data becomes available.

Derived attributes can be defined as functions depending on other nutrients and on the time. Regarding the time, data are sparse. This means that not all the needed information are available for a given time point, therefore derived nutrients cannot be always computed. We want to provide the value of a derived attribute through an estimation that uses nearest neighbor search whenever a needed value is not available for a given time point.

- #Vit.B = *f* (Vit. A, K) = Vit. A * K
- #OS = *f* (#Vit. B, Ca) = Vit. B / Ca

Sample	Feed	Time	Vit. A	Ca	K	#Vit. B	#OS
1234	Corn	2005	1.2		2.0	*2.4*	*2.4 / 4.7*
1235	Hay	2006	3.2		6.0	*19.2*	*19.2 / 3.9*
1236	Hay	2008	3.0	3.9		*18*	*18 / 3.9*
1237	Milk	2008	5.3		*4.0*	*21.2*	*X*
1238	Corn	2008		4.7	3.0	*9.9*	*9.9 / 4.7*
1239	Corn	2009	3.3		2.5	*8.25*	*8.25 / 4.7*

Nearest neighbor search will be applied on a non-fixed subset of the data source that changes based on the user selections. This is why derived attributes cannot be precomputed. Because the data set is chosen through a user-defined query, and due to the large amount of data, the solution must scale very well in order to be executed on the fly.

This work aims to get the knowledge of one approach to compute KNN-join, implement it, and show when it can be efficiently applied to the derived attribute computation in the Swiss Feed Data Warehouse.

Tasks:

1. Get familiar with the Swiss Feed Data Warehouse: read the schema and install the latest version of the Warehouse on the own machine.
2. Study the problem of computing derived attributes on the SFDW by reading the material provided by the supervisor and the introduction of 'The k-Nearest Neighbor Join: Turbo Charging the KDD Process'.
3. Understand the papers 'K Nearest Neighbor Queries and KNN-Joins in Large Relational Databases (Almost) for Free' and apply the proposed solution to the SFDW.
4. Produce a report listing where and why the proposed solution fails and/or fits for our study case.
5. Final oral exam (30 minutes).

Literature:

- **K Nearest Neighbor Queries and KNN-Joins in Large Relational Databases (Almost) for Free.**
  Bin Yao et al. ICDE 2010
- The k-Nearest Neighbor Join: Turbo Charging the KDD Process.
  Christian Böhm, et al. Journal Knowledge and Information Systems, 2003 *An introduction to KNN-Join with the SQL-solution*
- SW-Store: a vertically partitioned DBMS for Semantic Web data management D.J. Abadi et al. The VLDB Journal, 2009

A reference for a detailed understanding of the SFDW model's design

Supervisor:

- Francesco Cafagna

Starting date: 08/03/2012

Ending date: To be determined → 5/6/12

Department of Informatics, University of Zurich

Prof. Dr. Michael Böhlen