To: Prof. Rafal A. Angryk  
From: Atanu Roy  
Chandrima Sarkar  
Date: 16th Oct, 2009  
Subject: Research proposal

Proposed Research Topic:  
Fuzzy Querying Indexing over Imprecise Data with Taxonomies

Purposes:  
Imprecise data handling has been an increasing concern over the past few years. Several papers have also focused over how to query or index imprecise data with ease and increasing speed. But none of the research papers have delved into the process of taxonomically retrieving imprecise data. [1], [2], [3] have already worked on topics closely related to this field. But none of the researchers tackled the entire problem that we propose to tackle. In this paper we will try to implement an algorithm based on taxonomy or hierarchically grouping data which will make the querying of imprecise data much faster and easier. This in turn will aid in better searching, browsing and filtering the databases with imprecise data. If time permits we may also implement an indexing structure for efficient querying and searching.

Background:  
Though imprecise data is a common scenario in real world databases, finding a proper sample dataset for executing our algorithm will not be easy. Since large imprecise datasets are very hard to find we will use the existing datasets and synthesise data by changing the imprecision at different levels using existing data benchmarks. We will experimentally show the results of our techniques. This is because the available datasets are not large enough to provide concrete conclusions regarding the correctness of the algorithm we will propose.

Theoretical Framework:  
We will be guided mostly by [1] & [2]. We explore that organising imprecise databases hierarchically will aid better searching. We will propose in our research that the indexing structure we propose will work with hierarchical data. Here in this paper we propose to work on different kinds of taxonomies. A depth based tree with low branching and a breadth based tree with very high branching. For the depth based trees we may use a numerical dataset and for the tree with high branching we may use a dataset about a religion (categorical data). Since there are many religions in this world but religions rarely have more than few subdivisions (sects) inside it, thus the depth of a religion dataset cannot be very high, typically of the order 3. We propose to assign probabilities for the occurrence of each tuple, which we can in turn call “Degree of Relevance”. At the end of the research we would also like to show that our querying/indexing techniques will work with both kinds of taxonomies efficiently.

Method:  
The main methodology which we propose to implement is at first to devise an algorithm (inspired by the Naive Bayes classifier [5] or Support Vector Machines [7]) which can speed up... 

Chandrima Sarkar  
Atanu Roy
the classification of imprecise data into hierarchy. We will try to build the hierarchy of imprecise data based on both depth based tree with low branching and a breadth based tree with very high branching using Weka. After this text classification is done, we will implement fuzzy query [6] on this taxonomy of data hierarchy which should make data retrieval faster.

Timetable:
Prepare proposal by 16th Oct
Complete literature review by 20th Oct
Complete analysis by 20th Nov
Give presentation on 7th Dec - 11th Dec
Complete final report by 11th Dec (Not finalized by Dr. Angryk)

Limitations:
Time constraints (8 weeks) of the semester provide us with less time than may be ideal for this research. Since standardized imprecise datasets are very hard to find, therefore we may end up synthesising our own data to run tests on our algorithm. There may not be any available data benchmarks for imprecise data which fits our purpose. Thus we may have to create our own benchmarks which are a very tough job considering we have only 8 weeks to complete the research.

References:
2 - Indexing Uncertain Data – Pankaj K. Agarwal, Siu-Wing Cheng, Yufei Tao, Ke Yi.
3 - Scalable feature selection, classification and signature generation for organizing large text databases into hierarchical topic taxonomies - Soumen Chakrabarti, Byron Dom, Rakesh Agrawal, Prabhakar Raghavan
4 - Probabilistic Databases: Diamonds in the Dirt – Nilesh Dalvi, Christopher Re, Dan Suciu
5 - Improving Multiclass Text Classification with the Support Vector Machine - Jason D. M. Rennie, Ryan Rifkin
6 - Using a Fuzzy Classification Query Language for Customer Relationship Management - Andreas Meier Nicolas Were Martin Albrecht Miliadis Sarakinos

I'm not sure what you propose in "Method" section here. Was that what presented at your talk? Why are you guys going away from the original idea?

Chandrima Sarkar
Atanu Roy