Data Visualization Final Project Proposal

Adwin Jahn, Harley Montgomery

In the paper of Building Hierarchies of Concepts via Crowdsourcing, the machine learning system can actively select questions to ask using an information gain criterion. With the answers collected via crowdsourcing, the system can builds high quality hierarchies. In contrast, the experts often design one single hierarchy to best explain the semantic relationships among the concepts, and ignore the natural uncertainty that may exist in the process. For example, does tomato belong to vegetable or fruits, there would just be a single answer from one expert, but we want a probability over different hierarchies, such as 70% people believe that tomato is a fruit, and 30% things the hierarchy is different. Through more and more questions actively generated by system and answer collected, the hierarchy will be more and more accurate. With the accurate hierarchy of concepts, machine can know human better and solve problem by asking questions as children. The applications are in broad area including online shopping, customer service, and Artificial Intelligence. Demo video shows the evolution of hierarchy of concepts:
https://www.youtube.com/watch?v=tD4BSF8tIDQ&index=15&list=PLMdjyI1iH4ysIWKv82RfdD7nF5a_BQ0l6

There are several problems need to be addressed:

1) In video, every frame just shows the MAP hierarchy (tree), however, the ideal visualization should show some hidden edges, which are with high probabilities. Taking the concept ``tomato'' as an example, even though the MAP hierarchy shows that ``tomato'' is the child of ``fruit'', the weight of the edge between ``tomato'' and ``vegetable'' is also high. The high edge weight suggests that the uncertainty (whether the ``tomato'' belongs to ``fruit'') is high. To ashow the uncertainty, we hope to show the hidden edge when people hover over the hierarchy/tree nodes.

2)The hierarchy/tree of the current frame should not differ a lot from the previous frame. But many visualization tools do not consider the similarity constraint between consecutive frames, and would layout the tree to optimize the space usage of the tree layout. However, for a better visualization of the sequence of frames of hierarchy/tree, we need to have fixed tree render algorithm to minimize the difference between sequences
of frames.

3)Now the training data is collected by crowdsourcing. Crowdsourcing workers just answer the questions like questionnaires. However, we want to build an interactive interface that the front-end would show the current status of the hierarchy/tree. Such that workers would see how the hierarchy/tree changes when they feedback the answer. Instead of randomly choosing questions, the system deploys an active learning algorithm to choose the question, which is the most informative for the current status, in other words, the system keeps asking the question based on the previous feedback answered by users. More interestingly, we want to compare if people would have better correct answer rate when we show the status of hierarchies when they do the

questionnaires.