Loc Do (halocd)
Joseph Richardson (jmrichar)

Project Proposal

For our semester project, we will be implementing a recommendation system that utilizes text analysis and web technologies. In order to limit our scope, we have opted to make our system focus only on YouTube rather than on the entire web. Our aim is to organize videos into a system of categories so that, from any starting video, we can recommend a non-terminating sequential list of related videos that "walks" along so that, while any two consecutive videos are related, the sequence may wander from the original topic. Users are free to skip videos, and can adjust the degree of "curiosity", which determines whether the sequence is more likely to explore a variety of categories or stick to one or a small few.

In short, we intend to make a program that will allow users to experience YouTube in a manner akin to Pandora radio, where the program rather than the user selects individual videos to show. This differs from the current form of YouTube, where after each video a user must carefully consider all recommendations and choose one.

With this project, we hope to answer the question of "can the videos on YouTube be sufficiently categorized to allow non-human-guided selection of videos". We hope that the result will be directly practical as well for entertainment purposes.

The data for this project will be obtained via a web crawler we will write. Fortunately, YouTube sends its recommendations directly in the HTML result of an HTTP request of a video, so we can easily use those as the starting point for building our recommendation system. The data we will gather for each video will include its metadata (title, description, length, uploader, published date, number of views/likes/dislikes, user reviews, channel), and related (YouTube-recommended) videos. Our data will be converted into a directional graph before our ML program operates on it to create a categorization tree. Recommendations will then be determined based on this tree to respond to user queries.

Since the very final measure of of our performance will be user satisfaction, we plan to conduct a user study to evaluate our system. One possible study is to conduct an "A/B test", where users are provided with different recommendations from different approaches (including ours) and vote which one is the best. Currently we are looking for some existing research work related to our problem, and would report on them in the mid-semester report. Besides this, another performance measure can be to have human reviewers take a random sampling of videos and assess the subjective "relatedness" of those videos that our program considers to be in the same category.

Our project involves several stages. The first stage is to collect data, including building the YouTube crawler and setting up a data-storage system (where, due to the large amount of videos, scalability is a main concern). Concurrently with this, we would look for existing alternatives and published techniques on the video-recommendation problem. In the implementation phase, we would design both the main ML program and the UI (likely a web page), which both allows users access and allows us to collect feedback. The last phase is actually gather and analyze user feedback and to incorporate any apparent changes.

We define two types of goals: Our short-term goal (withing the scope of a course project) is to have a working recommendation app that, according to user feedback, 1) recommends a next video similar to the current one, 2) walks from category to category over time, and 3) is subjectively considered useful by a good number of users—but limited to some sub-set of videos on Youtube. Our stretch goal is to both cover more videos and to leverage external content (such as Wikipedia) to gather more information based on video metadata and improve our recommendations.

While extremely accurate recommendations may be difficult to achieve, we hope to progress towards it, and the educational value of this project will be very high regardless.