Information Retrieval - Project Plan

Rodrigo Doria Medina - Thibault Roucou

November 10, 2011

1 Description of the project

Our project is a search engine specialized in **recent information** using a **newspaper style presentation**. This search engine will retrieve articles, images, videos and comments. To find our documents, we will use different sources :

- For articles : Google News
- For videos : YouTube
- For images : Flickr
- For comments : Twitter and Google+

The sources of documents will be Google News, Twitter, Google+, YouTube and Flickr. We will also use Twitter and Google+ to expand the user query. The most ”popular” post will be used to search in the different engines. Using Social networks will allow us to find information that really interest people with the use of the public streams. We will also be able to give information which will be relevant for the user connected on the system using a ranking system based on trust in the people who posted the documents. We will also use built-in ranking system of each system, like for instance, the most recent and most viewed for the YouTube videos. But we will try to provide, when it's possible, very recent documents.

As a source, Twitter and Google+ will only be used for comments about the topics but we will not use the links provided in the posts.

Our ranking system will be used in the presentation of the results, the top ranked articles will be ”headlines” of our newspaper whereas the last relevant results will be in the ”brief” section of the page, with only small paragraph referring to it. For each headline, we will try to provide an article, one or more videos and one or more images. We will also provides post from twitter and Google+ as comments of the documents.

Social networks will also help us to find the best verticals for each query. If a query return 80% of videos on Twitter and Google+, we will give a better rank on videos found on YouTube than on images found on Flickr for example.

2 Evaluation of the results

To evaluate the results, we only can use the user feedback because the results will be specific to her. So we will put a button next to each result so the user can tell herself if it is relevant or not. In this case we will be able to evaluate the system. This evaluation system can also be used to improve the result of the results provided to the user.

3 Scenario

When a user access to the system, he can either choose one of the trending topics displayed from twitter and Google News or type her custom query in a searchbox. If she wants, she will be able to link her twitter and/or Google+ account in order to get personalized results.

The system will then process the query, and display the results in a newspaper like presentation. The user will therefore be able to play the videos, see the images, and read the articles and comments provided. For each document she will be able to express her opinion on the document by clicking on a button to say if the document is relevant or not for her.

Each time a document is said to be irrelevant, it will be replaced by another one from the next results.

4 Processing the query

The internal system will work as follow :

4.1 Receiving the query

The query is entered by the user and sent to the system. The system will receive it and send it to Twitter and Google+. The results retrieved by the two API will be sorted and the system will extract the important keywords to expand the query. If the user is logged in on Twitter and/or Google+ the system will use relations between the users to define which post are the most relevant for the user.

The most popular results from Twitter and Google+ will be stored for later use as comments.

4.2 Querying the sources

The query expanded will be sent to different APIs (Flickr, YouTube and GoogleNews) in order to get the documents. The aim is off course to find relevant documents but the system will try to focus on the more recent documents. The ranking system will therefore use a ranking system combining relevance and recentness.

The system will also use frequency of medias presents in the results from Twitter and Google+ to choose the percentage of medias to display in the final results compared to text. Indeed, if a query display 80% of videos in Twitter, it means that the users probably wanted videos and in this case, we will promote results from YouTube.

4.3 Displaying the results

After having ranked all the results, the system will display them to the user in a newspaper presentation with top ranked documents as headlines.

4.4 Receiving feedback

The user will finally express some feedback by stating if a document is relevant or not. The system will stored all these feedbacks in order to be evaluated later. It can also be possible to use these feedback to improve the ranking system for this user, next time she use the system.