Development of an Aggregated News Search Engine

Rodrigo Doria Medina and Thibault Roucou
Master Human Media Interaction
University of Twente, Netherlands

December 17, 2011

Abstract

1 Overview of the project

1.1 Description

The aim of the project is to create a search engine doing aggregative search by retrieving texts, images and videos using Twitter to build the ranking algorithm. Due to Twitter API restrictions, we decided to focus on news not older than 7 days. The results will be displayed as a NewsPaper.

1.2 Sources

1.3 Technologies used

1.4 Evaluation methods

2 Building the baseline

2.1 The APIs

2.1.1 Twitter

2.1.2 Google News

2.1.3 YouTube

2.1.4 Flickr

2.2 Choosing the documents for the baseline

3 The ranking algorithm

3.1 Passive ranking

3.2 Active ranking

3.2.1 Using popular words from Twitter

because of the system use with twitter, we obtain words that are useless for the query. Like for example when the user search for "Obama", the results which contains "president" or "barrack" will have a good score. But it's not giving any information about the news related to "Obama". SO it's better if the user enter "president barrack obama" as a query, the results are significantly improve in this case.

On the other hand, if the query is too precise, some providers like Youtube or Flickr will not return any results because they don't contain that much informations in their metadata.