## Assignment 1

**1.** Document filtering is an application that stores a large number of queries or user profiles and compares these profiles to every incoming document on a feed. Documents that are sufficiently similar to the profile are forwarded to that person via email or some other mechanism.

Document filters plays crucial role to provide efficient retrieval of document in the context of request. An intelligent filtering is needed in document detection applications, where millions of documents are received by organizations, while a relatively small subset of documents is of potential interest to any individual user.

In the scenario, filtering of document is required for every incoming document and to compare them against already existing profiles some techniques are implemented to match the individual user interest. For Example, Organization like dominos, papa johns stored user profiles which indicate points or rewards earned by any user. If the company wants to provide special offer to users with more than 100 points, offers are send to only those users with 100 or more points.

Architectural components for document filtering:

1. Influx of Documents

2. Indexing

3. Ranking

4. Evaluation

5. User Interaction

The Documents like Emails, web pages, newsletters, memos are influxes the data store. These documents are first Index to categorize for deciding any user context.

Index Creation combination of following methods:

1. Text Acquisition: It is implemented using techniques like crawling of web pages. This crawl web pages and follows links. Feeds which provide real time documents. It also employs number conversion techniques to convert document into consistent text plus metadata.

2. Text Transformation: Text Acquisition is followed by Text Transformation. This involves technique to convert sequence of document into tokens to recognize structural elements. This parser uses document syntax to identify structure. Another methods used while transformation is Stopping which removes common words and Stemming which usually group words derived from common root word. Link Analysis is used to gather information about the popularity and community of the linked pages. Classifier is used to identify the class related metadata.

3. Index Creation: It uses document statistics which gathers count and positions of the words and other features. For each index term frequency- inverse document frequency is calculated to provide the weight for each term. Inversion technique is used to convert document term information to term document information. For large number of documents, index distribution indexes across multiple computers.

Following the index creation Ranking is done for the document. Ranking is done to provide performance optimization and to provide efficient results in the distributed environment. This involves following methods:

1. Scoring: using ranking algorithm scoring of document is calculated.

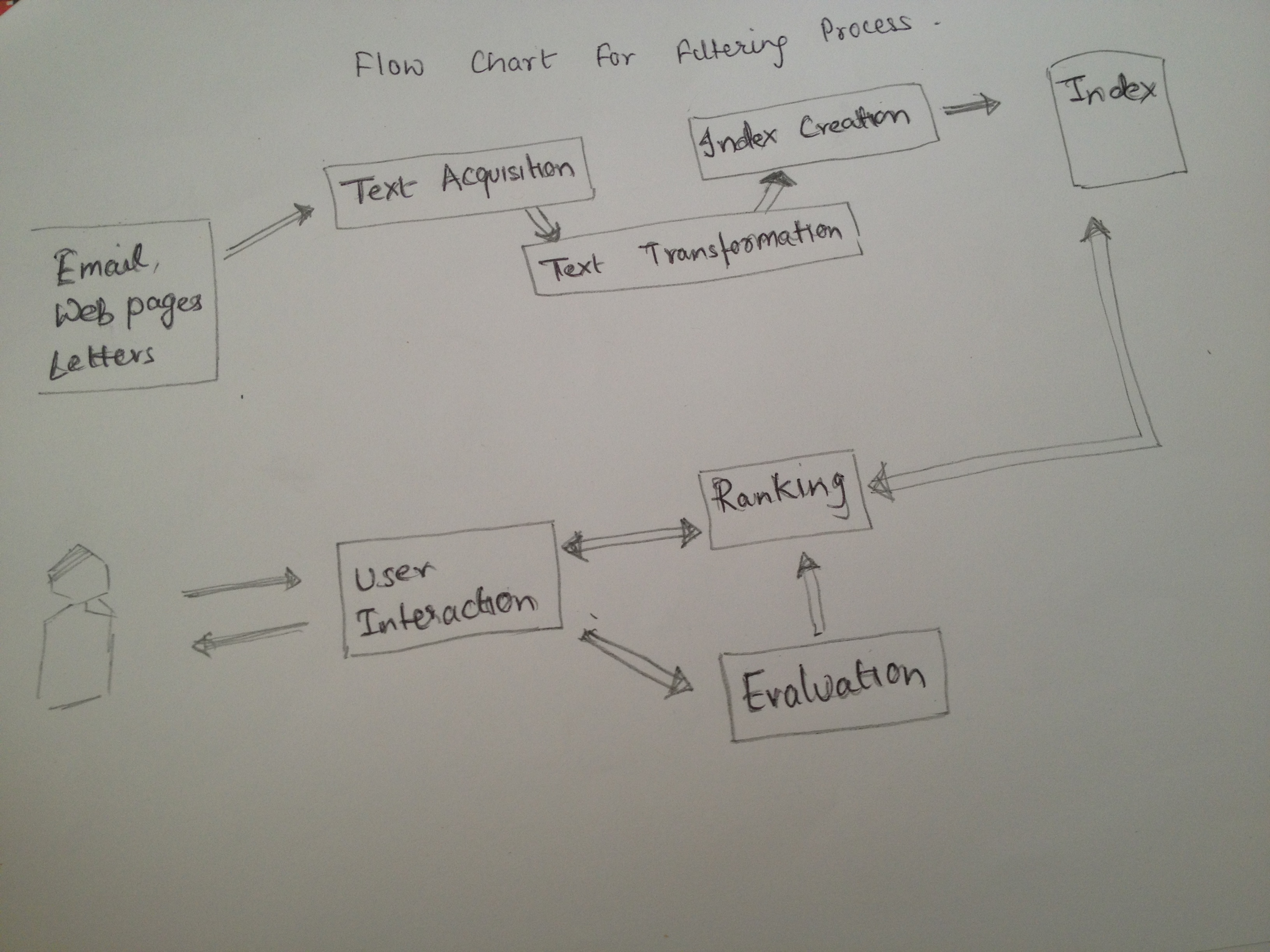
Following ranking document is evaluated by doing Logging Analysis which improves effectiveness of search and facilitates query logs, spell check. To measure system efficiency and ranking effectiveness, ranking analysis and Performance analysis is done.

User Interaction: User interaction is most crucial part of the filtering process. Based on the user input and feedback documents are processed to them. For each user query following methods are used to understand the context and interest.

Query Transformation: Improves the query before and after the search. Query expansion and relevance feedback modify the original term. Spell checking and query suggestion is used to improve query search.

Resultant Output of the query is the ordered documents according to the rank, it retrieves appropriate advertisement. According to the User input context is formulated and relevant documents are provided to the user.

Flow Chart for Filtering Process:



B. Major differences compared to Search Engine are:

1. Search Engine primarily searches for the query input from the data store. However Document filtering it compares the document against the already existing one.

2. In Search Engine matching is done against user query and all related keyword document is fetched and presented to user. Along with the user interest is logged in for further query processing.

While in Document filtering, matching is done against profile already present in the profile data store. User profiles are already defined and then any incoming document is presented to user if it specific to that user.

3. An Example of search engine google, yahoo, Bing. An Example where Document filtering is used rss feed, whether forecast, Pandora.

4. In Search Engine keywords are compared to fetch documents. While in document filtering documents are compared to fetch user profiles.

**2**. Command used to crawl the website ccs.neu.edu using GNU wget utility tool:

wget –user-agent=Mozzila/5.0 --recursive –delete-after –append-output=logfile –wait=5 –accept=html,xhtml,htm,pdf <http://www.ccs.neu.edu>

Please refer to the attachment named logfile.txt for the generated log.

**3**. Please find the attached source code and log generated for question number 3.