Emon Datta  
Student ID: 914329504  
Faculty Advisor: Dragomir Radev

CPSC 490 Project Proposal: Search Engine Summarization and Clustering

Background
Search engines have revolutionized the process of finding information online. They are used billions of times per day by users all around the world to answer questions on topics ranging from rocket science and nuclear physics to sports news and homeless service information. However, usability studies have shown that--despite their significant progress--search engines still suffer from issues that prevent users from finding what they’re looking for in a timely manner. Particularly if there are a large number of results to sort through and users are not familiar with ways to optimize queries to find exactly what they’re looking for, navigating long results lists can be very time consuming.

Overview
To help cut down the time taken to navigate search results and more easily find information, this project aims to add summarization and clustering capabilities to a search engine. In particular, it will first involve adding these capabilities to the AAN’s existing search bar and resource suggestion service for new projects. The second part of the project will involve creating a plug-and-play search engine solution that has summarization and clustering capabilities that can be added to websites.

Deliverables
1. Add summarization and clustering capabilities to AAN’s existing search and resource-suggestion services
2. Create a search engine with summarization and clustering capabilities that can be embedded in websites

Project Description
Part 1: Add Summarization and Clustering to AAN
Currently, ANN has a functionality such that people can search for relevant resources by entering in a query/topic in a search bar. It has another service in which users can enter the projects they are working on with a title and an abstract, and AAN will then recommend resources based on this information. In both cases, recommended resources are generally formatted as a list of links to papers (and some other resource types) with their titles, authors,
dates, and topic numbers. While this is certainly helpful to researchers, it could be made even more useful through the addition of summarization and clustering features.

A. Clustering
Clustering search results by common paper themes and how they relate to the query/project could help users more quickly and easily find the resources which are most useful to them--or, get a good idea of the breadth of the existing resources without having to click through every returned resource individually.

B. Summarization
Summarization can be useful in several ways when users are looking through the returned resources. 1) Having summaries of clusters of documents would help users navigate to clusters which are relevant to them more easily, and focus their energy on those relevant resource sets. 2) Having a brief summary of the resource pop up in a hover window when the title is hovered over could save both time that would otherwise be spent browsing irrelevant resources and bandwidth spent downloading/opening the resource.

Part 2: Plug and Play Search Engine
The ultimate goal of this project is to create a plug-and-play search engine that can be used to help web surfers quickly find information on websites. I became particularly interested in this based on common usability issues on websites, particularly the websites for non-profit and other non-tech services, which are usually run by less technical people. While search engine solutions through services like Google can be integrated into a website or domain, the results returned are just a list of ranked single pages or documents. While helpful, this can still be difficult to navigate. Including clustering and summarization features could greatly decrease the time taken to look at different pages, as they will be organized to optimize quicker learning times. Particularly for users who have less computer familiarity--which comprises a good number of those who are seeking services from institutions like non-profits--this could be the difference between them finding the services they need or not.

As such, this second part of my project involves creating a plug-and-play search engine that can be added to websites similarly to how Google’s search bar can. However, this search engine will also include clustering and summarization features (similar to those described above to be added to the AAN system) to expedite the search process.

Technologies
The AAN system currently uses SOLR to support the search capabilities, so I plan to start by using SOLR as well.
Methodologies & Algorithms

I plan to try several different algorithms for summarization and clustering in order to select which seem most appropriate for our purposes. Below, I have included reference papers with some of the algorithms I plan to begin investigating and testing.

A. Summarization
   1) Cluster Summaries & Key Words
      Automatic summarization of search engine hit lists (Radev & Fan)
      Information fusion in the context of multi-document summarization (Barzilay et al.)
      Multi-document summarization by sentence extraction (Goldstein et al.)
      Query-sensitive mutual reinforcement chain and its application in query-oriented multi-document summarization (Wei et al.)

      2) Individual Resource Summaries/Snippets
      A system for query-specific document summarization (Varadarajan & Hristidis)
      Enhanced web document summarization using hyperlinks (Delort et al.)

B. Clustering
   A comparison of document clustering techniques (Steinbach et al.)
   Document clustering based on non-negative matrix factorization (Xu et al.)
   Scatter/gather: A cluster-based approach to browsing large document collections (Cutting et al.)
   Web document clustering: a feasibility demonstration (Zamir & Etzioni)
Works Cited


Xu, Wei, Xin Liu, and Yihong Gong. "Document clustering based on non-negative matrix

Zamir, Oren, and Oren Etzioni. "Web document clustering: A feasibility demonstration."