CS 8803 ADVANCED INTERNET APPLICATION DEVELOPMENT PROJECT PROPOSAL – CLUSTERING WEB SEARCH RESULTS

SUBMITTED BY: ASMITA BARVE AND RUCHEEK SANGANI

MOTIVATION AND OBJECTIVES:

The world wide web, with the astronomical amount of information it stomachs, behaves often like a multi-storied mall. You are looking for a yellow jacket (preferably GAP), you know it is there, but you simply cannot find it. Search engines have proven to be an elixir for this distress and have enticed quite an amount of research work over the past few years. As the size of the internet and the number of users grow exponentially, the going gets tougher. Even more challenging is the fact that though technology advances and its beneficiaries grow in numbers, there is no change whatsoever in their mindset. Simply put, the average user still looks at only the first ten responses of the search engine, tries another search in some time and another search engine in some more!

More often than not the problem lies neither in the content of the internet nor in the mechanism of the search engine, but the way in which the retrieved results are presented. Consider, for instance, the aforementioned example of a yellow jacket. Conventional engines will give a list of links to outlets which have yellow jackets to sell, interspersed with links to the Georgia Tech football team. We envisage a search engine which clusters the results of each search into relevant categories. What the user sees after entering “yellow jackets” is a bunch of folders; one which says ‘Shops’, another which says ‘GT football team’ and a third which says ‘insects’ and so on.

With an interest in the fields of data mining and artificial intelligence techniques, clustering emerged out as the most natural candidate solution meeting the above needs. Our approach involves collecting results from various search engines, clustering them and presenting them in a manner much more appealing to the user.

RELATED WORK

Literature:

Attempts to organize web search results have been on since a decade. Oren Zamir, et. al have introduced in their paper [2], word-intersection clustering and phrase-intersection clustering. They propose the use of Suffix tree clustering for detecting phrase similarities amongst documents retrieved. Hua-Jun Zeng et al have compared five regression algorithms to cluster search results namely Linear, Logistic, SV-L, SV-R and SV-S [3] using different ranking criteria. Drawbacks of traditional clustering algorithms like HAC and k-means (the time complexity of the former and clustering quality for the latter) have led to efforts by researchers to explore new techniques. As stated by George Karypis et al [7], bisecting k-means (a variant of k-means) captures best of both the worlds of hierarchical and partitional techniques and is efficient. We plan to explore this algorithm and evaluate its performance in the case of clustering web documents.

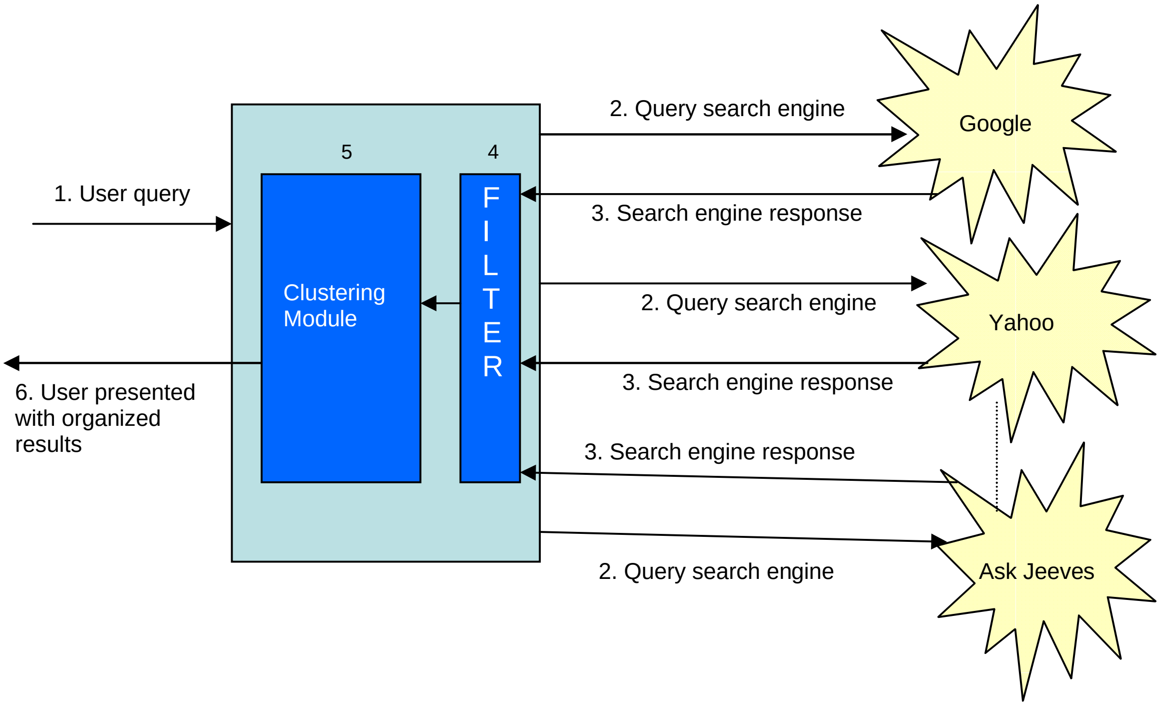
Systems:

Implementations of this idea of organizing search results have also come into being. ‘Clusty’ is a new meta-search engine made by Vivisimo. It uses all results provided by other search tools and clusters results so that you can refine your query by clicking on a more focused topic. Along with an interactive GUI it offers clustering by domain, topic and URL. Vivisimo also offers a similar service for corporate intranets where the makers of search engines are held responsible for both, the information content and the presentation of retrieval results. We want to extend the work done on these systems to include features that will represent the results in an even more comprehensive way.

PROPOSED WORK:

The basic idea of our work concerns with clustering the results obtained from various search engines. This idea is termed as Meta search engine, where we are building our own layer of search engine which will be an interface between user and other various search engines. This will provide us with the strong ability to manipulate on the results and allow us an opportunity to present them in a much more user friendly manner.

Figure below represents a simple **architectural design** for the project we wish to implement:



Explanation:

1. The users will be provided with an interface through which they can submit the query. This interface will also provide the ability to select search engines, type of clusters, clustering algorithms and so on. (More specified below)

2. This query submitted by the user on our interface will be fired to various search engines.

3. The search engine responds with the results that may include millions of web links pertaining to the query. However, these results are unformatted and random with respect to user requirements.

4. We will now scan through top ‘k’ responses (where k will be a parameter specified by the user) for additional keywords. The filter would also include an algorithm to avoid duplication of pages. Additionally, we plan to weigh the pages in these results in accordance to their rank in each search result.

5. These keywords and the documents will now be subjected to clustering algorithms. Past research have suggested the usage of ‘suffix tree’ algorithm and ‘bisecting k-means’ for clustering web documents. As we delve further into the project, we expect to learn about other clustering algorithms and use them too, to cluster the results. This will help us to compare the results of various clustering techniques and determine which one or which combinations provide pleasing outputs. We would also like to present the option of selecting clustering algorithm for the user. Clustering is kept as a separate module so that it can be easily plugged and replaced.

6. The result of clustering will now group the closely related documents into the same group. The documents will be basically grouped in accordance to the contents they contain. The clustered results will be shown in folders with all those pages belonging to one cluster falling into the same folder. The name of the folder too, will be dynamically created based on the cluster which would be representative for the cluster. The users based on their interest can promptly select the folder they were looking out for and browse through the folder containing data relevant to their query without worrying about non relevant pages occurring in their results.

Thus, the final aim of this project is subjective rather than objective with user satisfaction being the underlying motive.

Although work has already performed in this field, the most specific example being ‘Clusty [1]’ which is clustering based search engine (www.clusty.com), this project would be a good step towards improving our knowledge about the working of a search engine. Further, this project will enable us to compare effectiveness of various clustering algorithms on web documents. Finally, the user interface will provide various options to the user, some of which are not present in Clusty. These features include:

- Select which search engine/(s) to use.
- Number of top responses to be used for clustering.
- Clustering based on extension (eg: .pdf, .ppt.. useful for searching research papers)
- Clustering based on domain type (.com, .org, .edu .. useful for searching universities)
- Clustering based on server (eg: all pages belonging to cc.gatech.edu in same cluster)

PLAN OF ACTION:

Software: Java, open source clustering algorithms, client-side and server-side technologies for user interface
Hardware: Stand-alone computer
Operating system: Windows XP

Schedule:

Week No.	**Dates**	**Scheduled work**
Week 1	Feb 17- Feb 23	Study: Java n/w interface
Week 2	Feb 24 – Mar 2	Implementation: Extracting Results
Week 3	Mar 3 – Mar 9	
Week 4	Mar 10 – Mar 16	Study: Clustering algorithms
Week 5	Mar 17 – Mar 23	Implementation: Clustering Module
Week 6	Mar 24 – Mar 30	
Week 7	Mar 31 – Apr 6	Study & Implementation: The User interface
Week 8	Apr 7 – Apr 13	
Week 9	Apr 14 – Apr 20	Room for delays

We intend to make this project an insightful experience and aim to deliver a complete system keeping the ultimate goal of the web surfer's content in mind.

BIBLIOGRAPHY:

Following are few references that we have skimmed through and which we find to be a good support throughout our project duration.

1. Clusty – www.clusty.com
2. Oren Zamir, “Fast and Intuitive Clustering of Web Documents”, Quals Report, April 1997.
3. Hua-Jun Zeng, Qi-Cai He, Zheng Chen, Wei-Ying Ma, Jinwen Ma, “Learning to Cluster Web Search Results”.
4. Anton V. Leouski, W. Bruce Croft, “An Evaluation of Techniques for Clustering Search Results”, Technical Report IR-76, 1996.
5. Oren Zamir, Oren Etzioni, “Grouper: A Dynamic Clustering Interface to Web Search Results”.
6. Corporate Intranets, Vivisimo White Paper.
7. Michael Steinbach, George Karypis, Vipin Kumar, “A Comparison of Document Clustering Techniques”, Technical Report # 00-034.
8. Tagging versus Clustering for Enterprise Search Engines: Vivisimo White Paper # 3.