

Pre-Query Web Document Clustering

Adam Lewis

Undergraduate Masters Student

Lancaster University

a.lewis3@lancaster.ac.uk

Abstract

The areas of information search and retrieval are of great importance in modern computing/society. With an ever growing amount of new information being generated and stored daily, the need to enable users and provide flexible, powerful tools that utilise this information is becoming increasingly urgent. Current technology however lacks anywhere near the same levels of intuition and knowledge that a human can exhibit. This greatly restricts the capabilities of search engine users and cluster based search as a technology is believed to be capable of solving this issue.

Current cluster search technology is no where near perfect however and as such, this document argues in the favour of undertaking an exploration into cluster based search engine technology and performing experimentation on an innovative design conceived with clustered web search in mind. Executing such a project has the potential to help improve the current state of search engine technology and go on to inspiring those that will be building the next generation of search tools.

Categories and Subject Descriptors D.0 [General]: Exploration and experimentation with cluster based search engines

General Terms Documentation, Design, Experimentation, Theory

Keywords Initial Research Proposal; Dissertation;

1. Introduction

With the advent of the World Wide Web 21 years ago [1] and the global proliferation of computer technology, the topics of information search and retrieval have proved themselves to be key areas of interest for researchers. In response to the rapid growth of the web, a multitude of search engines have been founded [2], [3], [4] and have been proven to provide a very lucrative line of business when successful [5]. By responding to global demand, companies have been able to use catalogued information to their own benefit and can generate large amounts income from the traffic of their users. As holds true with the majority of the business world however, it is the companies that fail to innovate and remain contemporary that tend towards obsolescence and cease to remain profitable ventures. With this fact in mind, modern search companies are constantly trying to evolve and innovate to remain/become the market leader.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.

There is no better prior example of this occurring within the domain of search than the rise of Google as a provider. Before Google, search engines such as Altavista [6] and Lycos [7] were leaders and relied on a database of search terms to find pages that were closest compared to the user's query. "Unfortunately, this automated process did not always return the logical results. For instance, when searching for "Microsoft", pages for retailers selling Microsoft products might be featured before the Microsoft corporate homepage because a single page might list dozens of Microsoft products" [8].

It was at this time when the entire industry was focused on improving frequency powered search as described, it took just two students to come up with a revolutionary yet basic idea that would change the world and web search forever. By simply using the hyper-links on a web-page as "votes" [9] [25], web documents evolved into more than a simple collection of words and frequencies to be matched with a query, but became part of an ecosystem where pages could be ranked and ordered. This gave Google's founders an incredible advantage over competitors, with their PageRank algorithm outperforming the abilities of even the biggest search providers at the time.

The ability to provide a higher calibre of results allowed Google to defeat the competition with ease. By simply providing a superior service that was soon described as having "an uncanny knack for returning extremely relevant results" [12], Google established itself as the go-to search engine provider. Sixteen years later [10] and Google is one of the biggest names in the technology industry and dominates the competition in terms of traffic to its website [11].

This may not be the end of Google's story however as there are those that believe Google will eventually be out-innovated and will succumb to pressure from competitors as a new and better service presents itself to Internet users [13]. Attempts to defeat Google and provide a better service are many and range from changing the user interface [18], making search "personal" [19] and making it more intelligent [20]. These approaches are yet to achieve their goals, but one technological advancement that could potentially achieve this feat however is the development and proliferation of cluster based search technology.

2. Cluster Based Search

Cluster based search engines are not new in terms of the basic concepts required, and several examples exist on the web already that attempt to provide a viable solution for users [A]. As with the history of Google, these providers are all hoping to enjoy the same incredible success story and are generally much the same in most respects of the service they provide. For the purpose of this paper however, one example must be considered to effectively

demonstrate the differences between the current states of classical and clustered search. To do so, the reasonably well known search engine Clusty [14] was selected.

This choice was made not only due to Clusty's popularity however, but also for Clusty's design being largely similar to that of traditional search engines. This allowed for an increased focus on the functional differences between classical and clustered search instead of the presentational differences that can be found in engines such Mnemo [15].



Figure 1: Screenshot of Clusty's homepage

Figure 2: Screenshot of a search for “Apple”

As can be seen by figures 1 and 2 above, the design of Clusty is similar to any stereotypical modern search engine on the web. The home page provides a text field for search query entry, options to select what media type to search the Web for and a query submission button. Next, the results page provides a list of returned results, options to refine the query & advertisements in the form of sponsored results. These results are ranked by their relevance and everything thus far looks and behaves the same as any typical search engine.

Explore the options to the left however and Clusty's key advancement in search technology becomes apparent. When the user enters a query into Clusty, the tool does more than just return a list of results ordered by their importance – it clusters them into hierarchical groups & tags them. To see how this can be useful one need only consider a user searching for the term “Jaguar”. This user could be searching for the car, the operating system or the animal but without explicit instruction a traditional search engine returns results on all three.

With current technology, the only way to remedy this problem is to edit the query in the hope that by being more specific we can obtain the results we desire. Unfortunately however certain queries can be too ambiguous to easily qualify for this solution, such as the query “improvements made in search engine

technology”. It is when in such situations we can begin to appreciate the power of clustered search.

By allowing users to filter the list of results through the selection of relevant clusters, it becomes possible to provide a superior level of service. No longer would a user have to exert the additional mental effort and time required in order to compose a new query, instead they can just select from the scope of clusters within the results to locate the desired information. This topic of time wastage is a genuine concern on the web, with several examples of “counting seconds” research and users consistently displaying high expectations of a web service's behaviour [21].

3. Problem Statement

Although the concept and demand for a cluster based search engine has existed for quite some time, the technology hasn't quite caught on and isn't quite ready for wide spread usage. One of the key issues is the quality of clusters returned by search. Take for example a search for the term “Apple” on Clusty. One of the high level clusters returned is “Growing” which can be said to be quite accurate. Expansion of this topic however revealed the clusters “Technology”, “Fruit, Tree”, “Joyce Rosencrans, Post Food Editor” and “Other Topics”. This clearly undermines the validity of the clustering algorithm's results and when combined with the fact that successive queries will return entirely different result lists and clusters, the engine and its output can not be deemed to be anything but unreliable.

The subject of the results returned by Clusty also brings us onto another problem with modern cluster based search engines; The number of results returned by a clustering search engine is a greatly reduced subset of the total number of results available. A search on Clusty for example reduced the return set from 574,000,000 results to just 233 web documents. This is due to the way these engines are designed. When performing a query on a cluster based search engine such as Clusty, the process is designed to at first act like Google and simply find the set of results that matches the query (574,000,000 results). The service then takes a sub-set of the results and forms clusters with the documents that remain within the sub-set (233 results).

This cut down and unreliability is an unfortunate effect of building a cluster based search engine using current designs. As the clustering of documents creates overhead, performing clustering on the results of real-time searches is limited in terms of what it can achieve within a realistic time-frame. As discussed in “A Survey of Web Clustering Engines” [22], existing solutions handle this problem by restricting the dataset of a traditional query and then clustering the results before presentation to the user.

With this in mind, it is the goal of this project to attempt to solve the problem by proposing and testing the design of a system that performs the clustering of web documents during the crawling/storage process. I believe this to be a completely new approach due to the afore mentioned survey and with the most popular search results concerned about applying clustering to crawlers and improving their efficiency [23] [24].

Taking this approach innovates upon the design of current systems and we believe that building such a system will exhibit multiple advantages over current search-first cluster-later solutions. Primarily, we expect there will be a significant improvement in real-time performance through the removal of the clustering algorithm's overhead at query time. Secondly, we expect there will be an improvement in the number of documents that can be returned in response to a user's query as the system will not be limited to a sub-set only view. Finally we also expect that as a bi-product of increasing the number of documents

returned as clusters, the clusters themselves will become increasingly coherent and meaningful as more data becomes available to the cluster forming process.

4. Goals and Challenges

As stated by prior discussion, it is the aim of this research project to develop and test a cluster based search engine that undertakes the task of clustering at the time of crawling as opposed to the time of querying. As such, it is necessary to apply foresight and attempt the prediction of the project's goals and challenges that will likely to occur over the course of time. In order to bring the project to fruition, it is logical to first assert the three following stages to the project's time-line: design, implementation and analysis.

Stage one of the project, design, is expected to be the most heavy in terms of the researching efforts required. The initial hurdle we must face is finding data to effectively model the World Wide Web. As crawling the web in its current form is not an option due to its sheer size [16], we either need to use an archive from when the web itself was smaller or find a suitable model that will display the same qualities as the web. Two such examples that we currently believe would be suitable for use as models are a copy of a large Intranet, or a download of Wikipedia's database [17].

Once a model is found and available for use, the second and possibly the most difficult challenge we face will be formation of the crawler's design. Completion of this task will first require research into existing clustering and NLP algorithms, then using the gained knowledge to design a system capable of matching our basic concept. With no extensive experience in either of these areas, and the complexity associated with the development of an algorithm capable of scaling & distribution, I believe this will be the most challenging section of the project to complete.

Once finalised however the second stage, implementation, should prove much easier. With a design ready, the only challenges we can necessarily predict will be the unforeseeable set-backs that can occur with any programming project. Finally once developed, the system will undergo the analysis stage to define the success of the project.

As previously discussed, we expect a cluster-first search-later engine to exhibit a specific set of advantages when compared to typical cluster based engines; Namely we expect an improvement in the number of results returned, an improvement in the quality of the clusters and an improvement in performance of real time search. In terms of validating these claims however we are limited by what we can realistically achieve.

Attempting to compare performance between the prototype and other engines for example would be unreasonable without a common test bed for the systems to run on. Without access to system statistics in an identical environment, we could not state with any certainty whether our approach was faster or not. Additionally, the prototype system will be no where near the same calibre in terms of code efficiency when compared to other publicly available alternatives. As providers will have had and used large amounts of time to optimise their engines, it is unlikely our prototype will be able to even compete without a large amount of effort. In summary, a comparison of engine performance would be incredibly difficult to achieve and as such this facet will not be focused upon as a marker of this project's success.

Again and for similar reasons, we are also unable to test whether our system produced more or less results in comparison to others. Due to dependence on the model we select and the capabilities of individual engines, it may or may not be possible to

cross compare the outputs and judge which engines are more successful at returning clustered results. Hence if it proves to be a viable property for exploration in the future and the time is available for pursuing this activity, we will complete a comparison of systems.

Finally, this brings us to the task of assessing the quality of the formed clusters. Based on the analysis section of multiple papers concerned with this area of research, we can state with certainty that a user study will be the correct method to use. At this point in time we can identify two potential paths that may be taken by the study. If the search engine can be set up to use an identical data-set as existing solutions, it will be possible to perform a blind cross comparison with a study group of users. If this is not possible however, a user study will have to be performed to quantify user opinions of the system.

With either approach, user study techniques such as questionnaires and task-timing will be suitable for use and depending on the number of participants available, a within/between groups design can hopefully be used. As the exact experimental design can not be defined at this point in time however, the full details of the experimental plan shall have to remain undefined until the project is past the initial design stage.

5. Workplan

With twenty seven weeks (including holidays) between the submission of this document and the final deadline, a workplan must be formed. Using the three previously defined project stages: design, implementation and analysis, the project will be defined as following.

The design stage will be completed by the end of week 11 and must be completed no later than the end of week 12. During this time, I will select a suitable data model and then move onto researching NLP and clustering algorithms up until week 10. This will enable me to gain the most out of having face-to-face project tutor contact during term time. I expect however that this week 10 deadline will not be concrete and will likely overlap the time spent during the Michaelmas vacation creating potential system designs for proposal to/reviewing by my tutors in week 11/12.

Next, the implementation stage should prove relatively straightforward and has an optimum completion target of week 20. The system will be programmed in Java for the rapid prototyping ability it provides and the enormous collection of libraries/support available online when compared to other available languages. If the target of week 20 can not be met however, the target can be pushed back until the end of the Easter break for reasons discussed later. In terms of sub-targets however, it is not possible at this stage to define.

As a result, a redefining of the project's goals and deadlines will be set after completion and approval of the system's design to ensure a strong project work-flow is maintained. The success of this approach should make for a good topic of discussion in the week 15 intermediary results presentation however, and hopefully some form of progress will be available for demonstration by this point.

Finally the analysis stage will start at the end of week 20 providing the implementation stage does not over run. If this possibility is avoided as hoped, the four weeks of Easter break will be spent starting the full write-up and forming the experimental design. Once back at university for the summer term, the experiment(s) will be carried out, the results analysed and the paper finished.

6. Conclusion

In conclusion, this document proposes the construction of a cluster based search engine that performs clustering of documents at crawl time. Current technology is flawed and without research into the area these flaws will unfortunately remain. To do so will require innovation, research and actuation of our ideas. There are several challenges however that must be faced before realising this goal; Acquiring suitable data, designing the system and choosing how much to code to create/reuse are three such examples. We remain faithful that the project is plausible to complete within the given time-frame and will prove beneficial to this area of interest however.

Appendix

[A] List of found cluster based search engines:

- <http://search.yahoo.com/> Select “Directory” option, provides list based results with cluster based filters
- <http://clusty.com/> provides list based results with cluster based filters
- <http://www.mnemo.org/> provides a graph of clusters for filtering
- <http://www.polymeta.com/> provides list based results with cluster based filters
- <http://search.carrot2.org/stable/search> provides list based results with cluster based filters
- <http://search.yippy.com/> provides list based results with cluster based filters

References

All websites last known as active 16th November 2012

- [1] <http://www.wired.com/geekdad/2011/08/world-wide-web-20-years/> *Story of the world's first website*
- [2] <http://www.google.co.uk/> *Example search engine*
- [3] <http://uk.yahoo.com/> *Example search engine*
- [4] <http://www.bing.com/> *Example search engine*
- [5] <http://searchenginewatch.com/article/2193073/Google-Reports-12.2-Billion-in-Q2-2012-Revenue-35-YoY-Growth> *Google's revenue history*
- [6] <http://www.altavista.com/> *Altavista*
- [7] <http://www.lycos.com/> *Lycos*
- [8] <http://lowendmac.com/orchard/09th/google-history.html> *Quote relating to frequency based search engines, second paragraph*
- [9] <http://www.youtube.com/watch?v=oaHwXqvbHsc> *A simply video explaining the history & workings of PageRank*
- [10] <http://www.google.com/about/company/history/> *Google's history as told by Google*
- [11] <http://searchenginewatch.com/article/2158888/Yahoo-Search-Share-Sinks-Google-Bing-Rise-in-February-2012> *Proof of Google's popularity*
- [12] <http://web.archive.org/web/19991001132301/http://www8.zdnet.com/pcmag/special/web100/search2.html> *Quote of Google's uncanny success rates*
- [13] <http://www.bloomseo.com/will-google-reign-forever-the-rise-and-fall-of-an-internet-king/> *The prophesied downfall of Google*
- [14] <http://www.clusty.com/> *Clusty search engine*
- [15] <http://www.mnemo.org/> *Mnemo search engine*
- [16] <http://www.worldwidewebsize.com/> *Stats on the size of the web*
- [17] http://en.wikipedia.org/wiki/Wikipedia:Database_download *Reference to instructions for downloading Wikipedia's database*
- [18] <http://www.managedq.com/> *Search engine with a very unique interface*
- [19] <http://www.dailymail.co.uk/news/article-2085006/Google-search-gets-personal-using-Plus-decide-results-see.html> *Google's personal approach to search*
- [20] <http://www.wolframalpha.com/> *Intelligent search*
- [21] <http://blog.kissmetrics.com/wp-content/uploads/2011/04/loading-time-lrg.jpg> *Infographic of “second counting” and its importance on the web*
- [22] <http://search.fub.it/claudio/pdf/CSUR09.pdf> *A survey of web clustering engines*
- [23] <http://www.personal.psu.edu/pum10/tois-tan.pdf> *Clustering-based Incremental Web Crawling*
- [24] <http://dl.acm.org/citation.cfm?id=1321516> *Designing clustering-based web crawling policies for search engine crawlers*
- [25] <http://www.emba.uvm.edu/~tlakoba/AppliedUGMath/DeeperInsidePageRank.pdf> *“Deeper Inside PageRank”*