Using Hierarchical Clustering and Summarisation Approaches for Web Retrieval: Glasgow at the TREC 2002 Interactive Track

Richard Osdin, Iadh Ounis and Ryen W. White

Department of Computing Science
University of Glasgow, Glasgow, G12 8QQ, Scotland
osdinra, ounis, ryen@dcs.gla.ac.uk

1. Introduction and Motivation

Current search engines are typified as having a lack of precision, coupled with an elongated ranked list style of result presentation. When combined, these factors make relevant data extraction increasingly complex. The main investigation of our participation in the Interactive Track of TREC 2002 is to assess the effectiveness of new visualisation techniques for displaying the results of search engines.

Our current system, provisionally named HuddleSearch, uses a newly developed clustering algorithm, which dynamically organises the relevant documents into a traversable hierarchy of general to more-specific cluster categories. We have extended our TREC-10 summarisation tool to also allow the summarisation of multiple documents; whereby a summary paints a caricature of the contents of a cluster, rather than an individual document, thus allowing the user to provisionally judge a cluster's relevance prior to viewing its contents. The interaction between the user and the system is further developed by the aid of an information visualisation tool. Our primary assumption is that the combination of both hierarchical clustering and summarisation tools will aid users in their interaction with the system in the Web context.

2. Systems

Our baseline system acts as a metasearch engine, providing a generic interface capable of displaying the results from any web search engine, simply by defining a wrapper specific to the baseline system. For the purposes of these experiments, we retrieve the results only from the provided Panoptic system[1], returning results solely from the .GOV domain. This system simply displays the title, a 200-character description and URL for each document, to provide the user with an element of familiarity with systems such as Google.

Our experimental system, HuddleSearch, extends the properties of our basic system, by enabling the user to find relevant documents quickly by means of navigating within a traversable hierarchy of clusters. When a user views a cluster title he or she gains an overview of the documents contained within it, and is then able to narrow in and view only the documents within a specified segmentation of choice, at a lower branch of the tree. In this way we address the problem of information overload; the user is able to reduce the relevant documents set by continually filtering out irrelevant documents in search of information satisfying their need.

Figure 1 shows the path between generality and specificity; where the retrieved set of documents contracts as the user progresses deeper into the cluster hierarchy. Unlike the flat clusters hierarchy of search engines like Vivisimo[2], WiseNut[3], or Grouper [2], HuddleSearch organises the clusters into a hierarchy, providing a better structure for the result set. Figure 2(a) presents a screenshot of the system when used as a metasearch engine on the Web. The clusters are shown as folders at the top of the interface. The title of the folder is indicative of the cluster content and the number on the folder represents cluster size.

We have complemented the conception of a cluster hierarchy by investigating the use of query-biased summarisation, previously explored by the Glasgow Information Retrieval Group, to provide short passages indicative of individual document content [1]. However, we have extended this practice to allow the summarisation of multiple documents. In a similar way to the previous work, the summaries are created on-the-fly at retrieval time, prior to the results page being displayed. Hereby, we introduce the creation of cluster summarisation, where groupings of significant sentences

[1] http://www.panopticsearch.com
[2] http://www.vivisimo.com
[3] http://www.wisenut.com

extracted from documents within a cluster are combined to produce a summary indicative of a cluster's content. The chosen sentences are ones that have a high degree of match with the user's query.

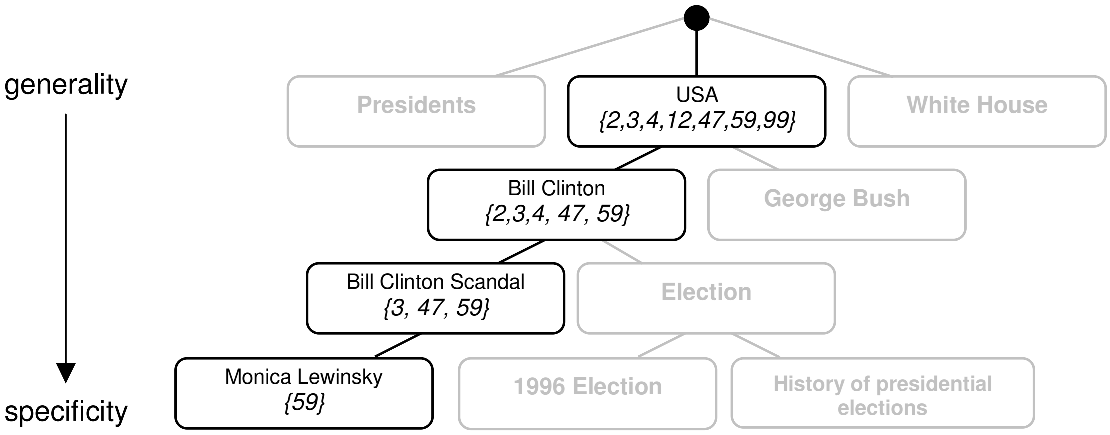
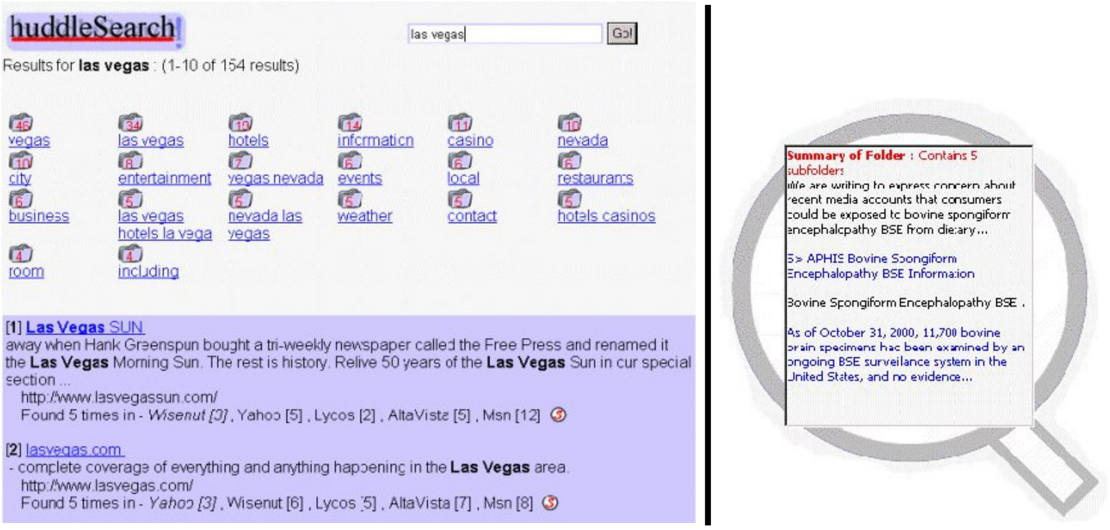


Figure 1: A dynamic hierarchical clustering approach

In addition to the traditional hyperlink method for navigating between clusters, as used by WiseNut, we have devised a visualisation tool, which allows the user to preview the contents of any given cluster with only the slightest mouse movement. This feature has been combined with our cluster summarisation; whereby a user can view the summary of any cluster with ease. The two complementary features enable the user to quickly glance at the contents of available clusters, initially assess relevance, and then select a cluster of interest. Searchers therefore do not waste valuable time viewing misleading document sets. Figure 2(b) displays our visualisation tool, which provides the user with a summary of a cluster's content when the mouse touches a cluster and appears on the display to the right of the clusters.



Results for **las vegas** : (1-10 of 154 results)

[1] **Las Vegas** SUN
away when Hank Greenspun bought a tri-weekly newspaper called the Free Press and renamed it the **Las Vegas** Morning Sun. The rest is history. Relive 50 years of the **Las Vegas** Sun in our special section ...
http://www.lasvegassun.com/
Found 5 times in - *Wisenut [3]* , Yahoo [5] , Lycos [2] , AltaVista [5] , Msn [12]

[2] lasvegas.com
- complete coverage of everything and anything happening in the **Las Vegas** area.
http://www.lasvegas.com/
Found 5 times in - *Yahoo [3]* , Wisenut [6] , Lycos [5] , AltaVista [7] , Msn [8]

Summary of Folder : Contains 5 subfolders
We are writing to express concern about recent media accounts that consumers could be exposed to bovine spongiform encephalopathy BSE from dietary...

5> APHIS Bovine Spongiform Encephalopathy BSE Information

Bovine Spongiform Encephalopathy BSE .

As of October 31, 2000, 11,700 bovine brain specimens had been examined by an ongoing BSE surveilance system in the United States, and no evidence...

Figure 2 (a): HuddleSearch interface (b): Cluster summary

3. Experiment

Two systems were used in our experiments: the provided Panoptic search engine, acting as a *baseline*, with its classical list-based approach, and the experimental system HuddleSearch, where the hierarchical clustering and the summarisation visualisation tool where activated. The HuddleSearch wrapper allows the Panoptic search engine results to be preserved[4], but masked the engine identity, hence avoiding any possible bias caused by previous searching experience. HuddleSearch and Panoptic were referred to only as System X and System Y respectively.

[4] Note that as mentioned in section 2, only the first 200-characters of each returned summary description is displayed.

A total of 16 users were recruited, each participant educated to at least a graduate level, from differing academic backgrounds, representative of the general web-using university populace.

Most users either work with or use computers for academic purposes frequently and have on average 5.7 years of web-searching experience. With the exception of just 1 user, Google was cited as being the search engine of choice.

Each user was required to carry out a total of 8 standard search tasks, equally split between the two systems. The tasks were allocated as required by the track guidelines to reduce potential learning effects and task bias. Figure 4 shows the tasks carried out by participants.

Following the track guidelines this year, each user was allowed a maximum of 10 minutes for each task. They were asked to use the system presented to them, and to perform the allocated task. All user actions were logged. Moreover, users were allowed to browse away from the result list to any degree. Due to the open nature of the collection this year, users were free to browse/save documents that were not in the TREC .GOV collection.

Government Regulation

- You are travelling from the Netherlands, and want to bring some typical food products as gifts for your friends. What are three kinds of food products from the Netherlands that you are not allowed to bring into the US? *(GR1)*
- You are concerned with privacy issues related to electronic information and would like to know what laws have been passed by the US Congress regarding these issues. Identify three such laws. *(GR2)*

Health or Project

- A friend has a private well which is the family's only source of drinking water. Locate a US publication, which contains guidelines for the maintenance of safe water standards for private well use. *(HP1)*
- You are not sure about the safety of genetically engineered foods, and would like to find more information and research on this topic. Name four potential types of safety problems that have been raised. *(HP2)*
- Name/find three research programs/projects that investigate the treatment/causes of dwarfism. *(HP3)*
- You are interested in learning more about what measures the US government has taken since 2001 to prevent Mad-Cow Disease. Identify three such measures. *(HP4)*

Travel

- You are planning a cycling expedition along the Silk Road in Central Asia. Find a website that is a good source information about health precautions should you take. *(T1)*
- You are planning to travel to the northeast territories of India and wonder if there are any problems/restrictions for tourists. Find a website that is a good source of information about such problems/restrictions. *(T2)*

Figure 4: Tasks used in TREC 2002 interactive track experiments

4. Results and Analysis

As mentioned before, all users actions were logged. Most of the data analysed in this section came from these logs generated by the system during the interaction with the users. All statistical tests of significance are at $p \leq 0.05$, unless otherwise stated. *M* is used in this section to denote the mean.

4.1 Task Completion

As part of the TREC post-task questionnaire, users were asked to state whether they felt they had successfully completed the task just attempted. We believe that ultimately, it is the user's decision to state whether he completed a particular task or not. Indeed, this reflects real-life situations, where the purpose of any system is ultimately to satisfy the user. Roughly speaking, our assumption is that if a user has stopped the task within the 10 minutes allocated time, and said it has been successfully completed, it means the user is satisfied and the task is marked as completed. Table 1 shows the total number of failures for each system (out of 64).

Table 1: Levels of task failure on each system

	Baseline	HuddleSearch
Total number of failures	15	9
Average number of failures	1.875	1.125

Table 1 shows that the number of incomplete tasks is *clearly* reduced by the use of the experimental system HuddleSearch. This shows that the clustering and summarisation features aid the users in their interaction with the system. However, paired *T*-tests revealed that the difference between the two systems was not significant (T_{14} = 1.43, p = .195).

4.2 Task Times

The times taken to complete tasks on both systems were automatically measured from the user logs. Table 2 provides an overview of the performance of both experimented systems. A 60 second penalty is added when the task is incomplete. The systems calibration times are taken into account in Table 2. Indeed, while the average number of submitted query per task on each system is almost the same[5], HuddleSearch is on average slower ($M_{\text{difference}}$ = 9.7 seconds) than Panoptic in returning documents[6].

Table 2: Average time per task (seconds)

	Task 1	Task 2	Task 3	Task 4	Task 5	Task 6	Task 7	Task 8
Baseline	589.5	645	376.13	406.88	417.88	416.63	489.13	218.5
HuddleSearch	544.8	394.93	294.3	309.05	271.8	463.05	461.18	225.55

Compared to the Panoptic search engine, used as a baseline in our experimental design, the times taken to complete search tasks using HuddleSearch have significantly fallen (see Tables 2 and 3), especially for the first 5 tasks. On average, 74.4 seconds are saved per task using the HuddleSearch system. This difference is more marked for some tasks than others.

Table 3: Average task completion time per system (seconds)

	Baseline	HuddleSearch
Average task completion time	444.9531	370.5813

Two-way repeated measures ANOVA was run to test for a link between system, task and the associated time for each type of task. The results of this test showed that there was a significant difference between systems ($F_{7,112}$ = 7.16, p = .001), and a significant difference between tasks ($F_{7,112}$ = 9.34, p = .000). We found firstly that hierarchical clustering and summarisation visualisation techniques significantly help the users to locate quickly the relevant documents and secondly, that not all tasks were of equal difficulty. If we assume that task completion time is a reasonable indicator of task difficulty, then Task 1 (GR1 in Figure 4) was significantly more difficult than any of the other seven tasks, across both systems.

4.3 User Satisfaction

Overall, 13 out of 16 users preferred HuddleSearch to the baseline. Furthermore, as part of the post-task questionnaire, users were asked whether they were satisfied with the search results for each task. Table 4 shows that users do feel more satisfied by the results provided by our hierarchical clustering system ($M_{\text{huddlesearch}}$ = 4.468 vs. M_{baseline} = 4.219).

[5] About 3 queries per task on each system.

[6] Distributed systems technologies are currently investigated to cut down the answering time of HuddleSearch, which is essentially due to the multiple documents summariser component.

Table 4: Average user satisfaction with results for each task (Scale 1 to 7, higher = better)

	Task 1	Task 2	Task 3	Task 4	Task 5	Task 6	Task 7	Task 8
Baseline	3.875	1.5	5	4	4.625	4.625	4.25	5.875
HuddleSearch	2.75	4.5	4.5	6	5.125	3.625	3.625	5.625

However, these results are not significant using two-way, repeated measures ANOVA testing the effects of system ($F_{7,112}$ = .432, p = .764) and task ($F_{7,112}$ = .453, p = .717) on user satisfaction. The degree of user satisfaction is not significantly different between systems.

Given the very encouraging results we obtained for the tasks completion, the average completion times, and user satisfaction, we wanted to investigate what could improve the overall effectiveness of HuddleSearch. We propose some possible explanations as to why the system was not optimally efficient.

Firstly, Panoptic, the baseline search engine returns a document result set consisting of documents fully matching the query, followed by a set of results only partially relevant to the query (results are presented in tiers, in Panoptic terminology). When a user makes a request for precise information and asks for 300 pages to be returned, Panoptic might simply return 3 fully relevant documents, which are then occluded by a further 297 only partially relevant and not fully satisfying the query. In this way, many of the base clusters created may be too generic, and relevant to only a subset of the query terms, thus not aiding the user in his search.

A cluster is fundamentally a common group of terms; this has proved to conflict with Panoptic, which often returns mirrors of the same document several times. At this moment in time, HuddleSearch simply examines the URL to determine replication; consequently 'false' significant clusters are created when content is repeated, this will result in incorrectly weighted common phrases and will hide naturally relevant clusters which are pushed further down the list.

In order for a cluster summary to be created each page within a cluster is visited and its content is retrieved to generate a page summary. Some experiments were executed at peak times, where network traffic was high; hence many documents failed to be summarised within an allocated time. As a result, when documents failed to be summarised, the higher-level cluster summaries were inadequate and occasionally users were faced with 'no summary could be created', providing little or no benefit to the user. However, informal feedback from users during non-peak time evaluations suggests that summaries are indeed helpful.

In addition to viewing a cluster summary, our visualisation tool provides the facility of previewing a cluster by looking at its top documents titles, which we believe to be good indicators for judging initial relevance. However, the .GOV collection made this feature extremely temperamental, as in many cases a document's title is simply an alphanumeric document ID, thus providing no hint of content.

The above perhaps explain why the user satisfaction with HuddleSearch, though superior to the Panoptic engine, was not optimal. Moreover, the assessment of the answers provided by the users during the experiments will provide more evidence regarding the effectiveness of HuddleSearch. However, we do believe that most of the issues mentioned above could be addressed efficiently. Hence, we suggest that there is much scope to improve HuddleSearch. Overall, results show that hierarchical clustering and summarisation visualisation tools do aid the users in their interaction with the search engine in the Web context. Uncompleted tasks as well as average times to accomplish them were definitely reduced by the use of HuddleSearch.

Acknowledgements

We would like to thank the Research Committee of the Department of Computing Science for financially supporting this work. We also like to thank those who participated in the above experiments; their comments, feedback and enthusiasm were greatly appreciated.

References

[1] White, R.W., Jose, J.M. and Ruthven, I. 'Comparing Implicit and Explicit Feedback Techniques for Web Retrieval: TREC-10 interactive track report'. Proceedings of the Text REtrieval Conference (TREC 2001). Gaithersburg, Maryland, USA. November 2001

[2] Zamir, O. and Etzioni, O. 'Grouper: a dynamic clustering interface to Web search results'. In Proceedings of the Eighth International World Wide Web Conference. Toronto, Canada. May 1999.