

CubanSea: Cluster-Based Visualization of Search Results

Matthias Tilsner, Orland Hoerber, and Adrian Fiech
Department of Computer Science
Memorial University of Newfoundland
St. John's, NL, Canada A1B 3X5
{matthias.tilsner, hoerber, afiech}@mun.ca

Abstract—In recent years, there has been a move toward supporting the human element of Web search beyond a simple query box and a ranked list of search results. In this paper, we present a prototype implementation of an approach to Web search based on fuzzy clustering and visualization. CubanSea presents the searcher with topics that are automatically discovered within the search space. A visual encoding of the cluster membership supports the searcher in understanding the relationship between the search results and the fuzzy clusters. Clusters can be selected for evaluation, and filtering operations allow the searcher to further manipulate and explore the search results set. The system is designed to support searchers when their information needs are ill-defined, vague, or ambiguous.

I. INTRODUCTION

The traditional approach to representing Web search results has been to provide all the matching documents in a list. An underlying assumption within this list-based representation is that the order implies the relevance of the document as determined by the search engine. When the information being sought is very specific, this list-based approach can be extremely effective, with many relevant documents often appearing at the top of the list. However, when the queries provided by the searchers are ill-defined, vague, or ambiguous, the list provides little support for the searcher to discover the few relevant documents from the many irrelevant documents.

The downside of this list-based representation is the limited amount of information that can be perceived and processed by the searcher. The only element that is available for encoding meta-data regarding the search results set (e.g., the ranked order) is the horizontal position of the search results within the list. Furthermore, search results lists are normally split into multiple pages containing a fixed number of results. This limits the ability of the searcher to make comparisons of search results to one another.

In this work, we attempt to address the shortcomings of the list-based representation, using fuzzy clustering techniques and visual approaches to search results representations. We do not try to improve or revise these techniques, but rather use existing solutions and concentrate on the visualization and interaction with the results. Due to the familiarity of the general public with list-based representations, we do not dispense with the list entirely. Our goal is

to augment the list with additional information that can help the searcher in those situations when the traditional ranked order of the search results does not provide adequate support for their search tasks (e.g., when the search topic is vague).

Our research has been guided by Shneiderman's visualization principle: "overview first, zoom and filter, then details on demand" [1]. This approach to interacting with the information available to achieve some task has been shown to be useful in numerous information visualization systems [2]. The application of this principle to the visual representation of Web search results is described in this paper, and implemented in the CubanSea prototype¹.

The remainder of this paper is organized as follows. Section 2 provides an overview of work related to this research. The details of the clustering and visualization techniques employed in CubanSea are provided in Section 3. The interactive support provided to the searcher as they seek relevant documents is discussed in Section 4. Conclusions and future research directions are outlined in Section 5.

II. RELATED WORK

A significant amount of research has already been undertaken in the area of search results representation. Most projects propose designs that position result items next to each other according to their similarity. LibViewer suggests a bookshelf-like design, creating different areas for different topics [3]. Result items are placed in accordance to their membership in these topics, using a self-organizing map. PEx-WEB proposes to position search results on a two-dimensional canvas, encoding the similarity of results in the distance between them [4]. A multi-dimensional vector is generated for each result, and projected onto a two-dimensional plane. DocuWorld uses a similar technique, projecting the vector into a three-dimensional space [5]. Each of these techniques requires the user to deduce the semantic meaning of the topics by examining the results they contain. Furthermore, they position documents that belong to more than one topic in the space between the topics. This, however, becomes hard to decode as the number of topics exceeds the number of dimensions available for positioning.

¹<http://cubansea.tilsner.eu/>

A number of papers suggest using color to encode the relevance a result has to a search query. Card Visualization splits the query into its different terms and assigns each a specific color [6]. An icon is presented for every result that encodes the relevance of the result to its most relevant term using color saturation. TileBars follows a similar technique, but uses vertical space to encode each group of query terms [7]. HotMap uses horizontal space for each query term, representing frequency of use with colours on a heat-scale [8].

Cluster-based approaches to Web search results representations are relatively common [9], [10], [11]. The goal of these techniques is to organize the search results into distinct groups, which can subsequently be explored by the searcher. While these interfaces can be quite useful, difficulties arise when documents describing multiple topics are placed in the cluster representing one topic, and therefore must be excluded from the other clusters to which it is also relevant.

III. CLUSTERING AND VISUALIZATION

The approach employed in CubanSea has been influenced by each of these systems described in the previous section. Rather than crisp clustering, a fuzzy clustering approach is employed to allow search results to be included in multiple topics simultaneously. The goal of is not only to organize the search results, but also to allow the searcher to gain some insight and understanding about the makeup and features of the search results space. A visual approach to representing the degree of membership in the clusters is employed, allowing the searcher to make sense of the clustering outcome.

CubanSea abbreviates “cluster-based visualization of search results”. Instead of providing a traditional ordered search result list, the CubanSea interface returns a set of distinguishable areas, each containing a subset of the total result list. These subsets overlap, covering the entire result list. Each area corresponds to one topic occurring in the result space. Figure 1 shows the visual representation of the search results for the query “piracy”. Two different topics have been automatically identified: “software report information” refers to software piracy, or piracy of digital media in general, while “pirates robbery reports” refers to sea robbery. These topic headers have been automatically generated and provide the viewer with the ability to immediately grasp the features of the topic space. Note that the clusters described in this example are created based upon the result snippets which are not displayed in the overview.

A. Search Results

The CubanSea interface focuses on the tasks of representing and visualizing the search results. The results themselves are retrieved from a search engine provider. For the purpose of this paper, the Yahoo Search API [12] has been chosen. However, the system is designed to work with any search

engine. In order to generate the clusters, a sufficiently large set of search results must be retrieved. Preliminary experimentation has found satisfactory clustering performance based on the top 50 search results.

B. Vector Model and Stemming

After retrieving the initial search results, it is necessary to convert these results into high-dimensional vectors that can be processed by the clustering algorithm. *Porter’s Stemming Algorithm* provides a convenient method for combining the various morphological variants of common base terms (stems) [13]. The occurrence frequency of these stems are counted, resulting in the value of the vector in the dimension corresponding to the unique stems.

The generation of the vector representation is based on the textual contents of the titles, snippets, and URLs, as provided by the search engine. While it would be possible to also retrieve the entire textual contents of each of the search results, this would introduce a significant delay between when the searcher submits their query and when the search results can be represented.

C. Fuzzy Clustering

The clustering algorithm chosen for this system is the *C-means fuzzy clustering algorithm* [14], [15]. This algorithm iteratively calculates a predefined number of centroids in a high-dimensional space and evaluates the membership of the individual items to these centroids using the formula in Equation 1. In this equation, u_{ij} is the membership value of search result i in cluster j . While v_i represents the high-dimensional vector of result i , c_j holds the vector representing the centroid of cluster j ($|c|$ denotes the number of clusters). The parameter m is a so called “fuzzifier” determining the crispness of the clusters. Values for this parameter between 1 and 2.5 have been shown to be effective [16]. Preliminary experiments have found a value of 1.5 to be a reasonable choice for our purpose.

$$u_{ij} = \frac{1}{\sum_{k=0}^{|c|} \left(\frac{\|v_i - c_j\|}{\|v_i - c_k\|} \right)^{\frac{2}{m-1}}} \quad (1)$$

The algorithm supports virtually any variation on a distance function for the calculation of $\|x_i - c_j\|$. Due to the non-normalized nature of the vector-based representations of the documents and cluster centroids, a cosine similarity function is used. Each iteration of the fuzzy clustering algorithm re-calculates the centroids of the clusters and then re-determines the fuzzy membership values for the documents. As soon as the re-calculation of the centroids does not yield a significant change, the algorithm terminates, returning the current centroids as the resulting centers of the clusters.

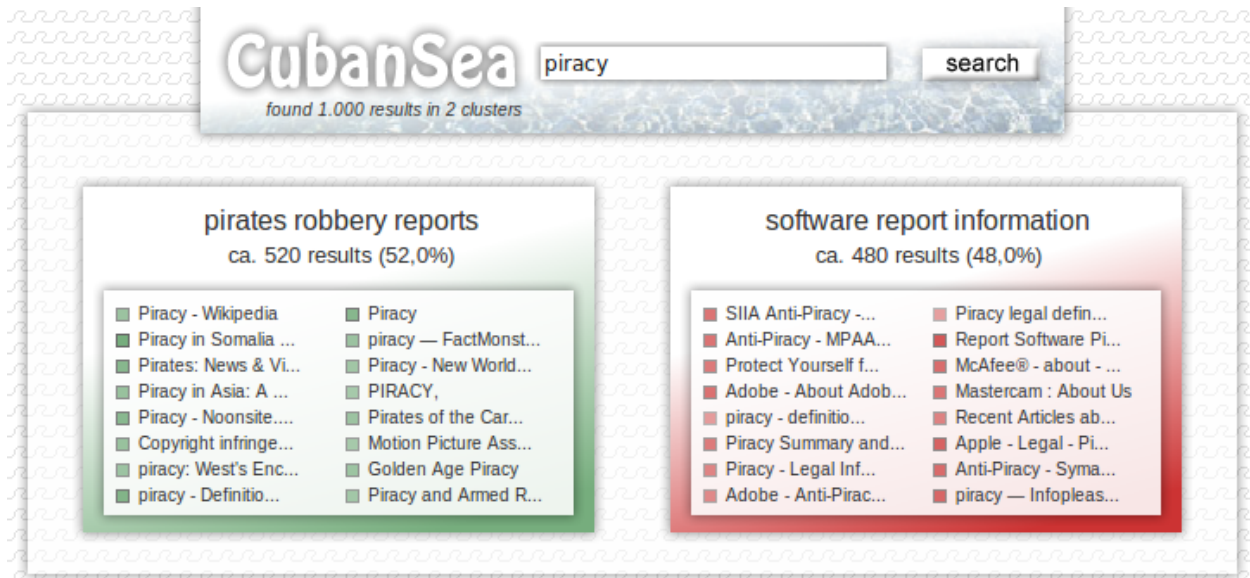


Figure 1. Clusters generated for the query “piracy”

An important element of the C-means fuzzy clustering algorithm is that the number of fuzzy clusters must be predetermined. CubanSea generates four clusters. This number is based on the assumption that a set of search results will seldom contain information on more than four distinct topics. However, in many cases, there will be fewer than four distinct topics within the search result set. In this situation, some of the fuzzy clusters may be very near one another. When two cluster centroids are determined to be very close, they are merged to produce a more meaningful cluster.

While each document will have some degree of membership within all of the clusters, the final membership for display purposes is based on the documents meeting a minimum similarity threshold with the cluster centroids. This allows individual documents to appear in multiple clusters, as was the goal with using a fuzzy clustering algorithm. In the case where a document doesn't meet the minimum similarity for any clusters, it is added to the single cluster to which it has the highest similarity.

The titles of the clusters are determined using the most frequent stems appearing in the topic. The actual term displayed is based on the most frequent terms within a given stem. The search terms are ignored in this process since they are likely to appear frequently in all clusters, providing little ability to distinguish between the clusters. While this process does not guarantee to generate unambiguous topic headers, it does provide results which are sufficient for this application.

Fuzzy clustering can be a performance bottleneck in many applications. In order to prevent this bottleneck, we chose to terminate the algorithm after a fixed number of cycles. Experiments have shown that the influence of this termination on the generated cluster space is neglectable. It

stands to reason that an increasing amount of algorithm cycles is a sign of an indistinct topic space which would in any event fail to generate highly meaningful clusters. The delay introduced by the fuzzy clustering algorithm is negligible in comparison to the delay in obtaining the search results set from the search engine provider.

D. Visual Encoding

With the goal of increasing the ability of the searcher to perceive the different topics, each is encoded with a distinct color. These colors have been chosen in accordance to the opponent process theory of colour [17], identifying green, red, blue, and yellow as being the most effective color set for labeling. Since yellow has a very low luminance contrast with the white background, it has been replaced with a dark grey. In order to prevent using high volume colors on large areas which would distract the viewers attention (as pointed out in [18]), a light gradient has been chosen. This gradient is sufficient for conveying a basic understanding of the cluster colors to the user without providing visual overload and drawing attention from the content. The top 20 results of each cluster are provided to aid the in topic recognition in the event that the headers might be ambiguous. The results are printed on a white-faded background to counter the interfering effect a gradient background would have on reading text placed on the foreground.

IV. INTERACTION

The methods for interacting with CubanSea, as searchers seek relevant documents, is guided by Shneiderman's principle of “overview first, zoom and filter, then details on demand” [1]. The discussion thus far has been focused on the “overview first” element.

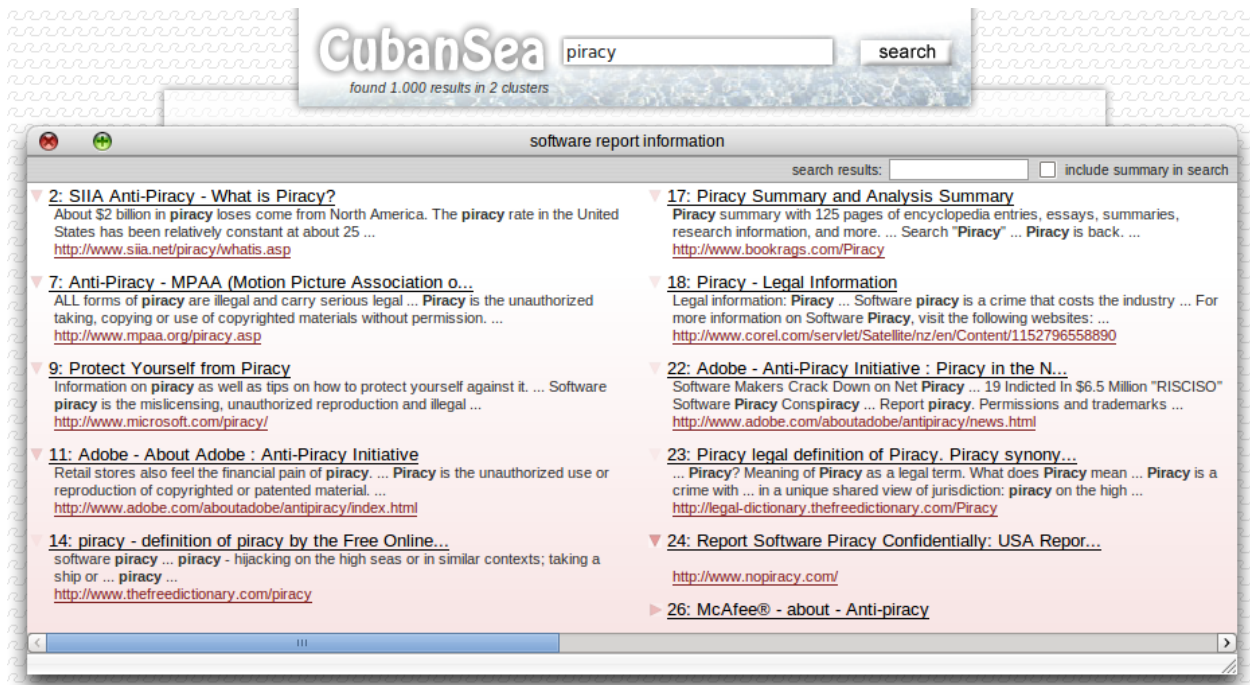


Figure 2. Displaying the “pirates history robbery” cluster

Once searchers identify a cluster of documents which appear to be relevant to their information seeking goals, they can click on it to “zoom” into this area of the search results space (see Figure 2). The ambient colour in the window is the same as that from the overview, providing visual feedback with respect to the selected cluster. Further, the intensity of the arrow icon beside each search result represents the degree of membership of the search result in the cluster. The user is also able to conduct a search within the search results set, implementing a “filter” operation. As terms are entered into the search field, only those search results that match the query are shown.

Within the search results list, only the first few search results are expanded, allowing the searcher to consider and evaluate these search results easily. The remainder are collapsed, allowing the searcher to see more of the search results set. If necessary, the searcher can view the “details on demand” by expanding a collapsed search results, or clicking on a search result to view the entire document.

The goal in following Shneiderman’s principle is to allow the searcher to first get a sense of the search results space, focus on a subset of the space that is relevant to their goals, and then access the details of potentially relevant documents as needed. While this may result in extra work for targeted searches where a single document is sufficient to address the information need, it can be very useful for vague or ambiguous searches where the documents are on multiple different topics. In the traditional list-based approach, the user must consider each of the search results one-by-one

for relevance. With CubanSea, the searcher can evaluate the clusters first, and select the one that most closely matches their goal for the search. The net effect is a pruning of the search results set, reducing the number of documents the searcher must consider.

V. CONCLUSIONS & FUTURE WORK

CubanSea provides a novel method for representing search results, providing an overview of the search results space, the ability to zoom to an area of interest and filter the search results, and allowing the searcher to access the details as needed. The system is designed to address the needs of searchers when their information seeking goals are ill-defined, vague, or ambiguous.

Within the context of Web information retrieval support systems [19], CubanSea provides support for searchers to find useful information and knowledge from Web resources [20]. Searchers are able to take an active role in the search process, making high-level selections of fuzzy clusters of documents, thereby reducing the number of non-relevant documents within the search results list.

Future work includes refining the interface and conducting laboratory experimentation to evaluate the value of the specific elements of CubanSea. Additional research is needed on the topic of performing fuzzy clustering based on the limited information available in Web search results.

REFERENCES

- [1] B. Shneiderman, “The eyes have it: A task by data type taxonomy for information visualizations,” in *Proceedings of*

- the 1996 IEEE Symposium on Visual Languages, 1996, p. 336.
- [2] B. Shneiderman and C. Plaisant, *Designing the User Interface*, 4th ed. Addison-Wesley, 2005.
- [3] A. Rauber and H. Bina, ““‘Andreas, Rauber’? Conference pages are over there, german documents on the lower left...”: An ‘old-fashioned’ approach to Web search results visualization,” in *Proceedings of the 11th International Workshop on Database and Expert Systems Applications*, 2000, pp. 615–619.
- [4] F. Paulovich, R. Pinho, C. Botha, A. Heijs, and R. Minghim, “PEX-WEB: Content-based visualization of Web search results,” in *Proceedings of the 12th International Conference on Information Visualisation*, 2008, pp. 208–214.
- [5] K. Einsfeld, S. Agne, M. Deller, A. Ebert, B. Klein, and C. Reuschling, “Dynamic visualization and navigation of semantic virtual environments,” in *Proceedings of the Tenth International Conference on Information Visualization*, 2006, pp. 569–574.
- [6] S. Mukherjea and Y. Hara, “Visualizing World-Wide Web search engine results,” in *Proceedings of the 1999 IEEE International Conference on Information Visualization*, 1999, pp. 400–405.
- [7] M. Hearst, “Tilebars: Visualization of term distribution information in full text information access,” in *Proceedings of the ACM Conference on Human Factors in Computing Systems*, 1995, pp. 59–66.
- [8] O. Hoerber and X. D. Yang, “The visual exploration of Web search results using HotMap,” in *Proceedings of the 2006 Conference on Information Visualization*, 2006, pp. 157–165.
- [9] O. Zamir and O. Etzioni, “Grouper: A dynamic clustering interface to Web search results,” in *Proceedings of the Eighth International World Wide Web Conference*, 1999, pp. 1361–1374.
- [10] Vivisimo, “Vivisimo search engine,” <http://www.vivisimo.com/>, april 23, 2009. [Online]. Available: <http://www.vivisimo.com/>
- [11] Grokker, “Grokker search engine,” <http://www.grokker.com/>, april 23, 2009. [Online]. Available: <http://www.grokker.com/>
- [12] Yahoo Inc, “Yahoo! search Web service,” <http://developer.yahoo.com/search/>, march 15, 2009.
- [13] M. F. Porter, “An algorithm for suffix stripping,” in *Readings in Information Retrieval*, K. S. Jones, Ed. San Francisco: Morgan Kaufmann, 1997, pp. 313–316.
- [14] J. C. Dunn, “Fuzzy relative of the isodata process and its use in detecting compact well-separated clusters,” *Journal of Cybernetics*, vol. 3, no. 3, pp. 32–57, 1973.
- [15] J. C. Bezdek, *Pattern Recognition with Fuzzy Objective Function Algorithms*. New York: Plenum Press, 1981.
- [16] C. Stutz, “Anwendungsspezifische fuzzy-clustermethoden,” Ph.D. dissertation, TU München, Sankt Augustin, 1999.
- [17] B. Berlin and P. Kay, *Basic Color Terms: Their Universality and Evolution*. Berkley: University of California Press, 1969.
- [18] C. Ware, *Information Visualization*. San Francisco: Morgan Kaufmann, 2004.
- [19] O. Hoerber, “Web information retrieval support systems: The future of web search,” in *Proceedings of the IEEE/WIC/ACM International Conference on Web Intelligence - Workshops (International Workshop on Web Information Retrieval Support Systems)*, 2008, pp. 29–32.
- [20] Y. Yao, “Information retrieval support systems,” in *Proceedings of the IEEE International Conference on Fuzzy Systems*, 2002, pp. 1092–1097.