Tag Clouds for Summarizing Web Search Results

Byron Y-L. Kuo[1] bkuo@mrl.ubc.ca
Thomas Hentrich[1, 2] Thomas.Hentrich@sfu.ca
Benjamin M. Good[1, 3] goodb@interchange.ubc.ca
Mark D. Wilkinson[1] mwilkinson@mrl.ubc.ca

[1] Affiliation: iCAPTURE Centre for Cardiovascular and Pulmonary Research. The University of British Columbia. St. Paul's Hospital. Vancouver, British Columbia, V6Z 1Y6, Canada. 1-604-2344 ext 62129
[2] Computing Science. Simon Fraser University. Burnaby, British Columbia, Canada
[3] Bioinformatics Graduate Program. The University of British Columbia. Vancouver, British Columbia, Canada

ABSTRACT
In this paper, we describe an application, PubCloud that uses tag clouds for the summarization of results from queries over the PubMed database of biomedical literature. PubCloud responds to queries of this database with tag clouds generated from words extracted from the abstracts returned by the query. The results of a user study comparing the PubCloud tag-cloud summarization of query results with the standard result list provided by PubMed indicated that the tag cloud interface is advantageous in presenting descriptive information and in reducing user frustration but that it is less effective at the task of enabling users to discover relations between concepts.

Categories and Subject Descriptors
H.3.3 [**Information Storage and Retrieval**]: Content Analysis and Indexing – *indexing methods, linguistic processing* .

General Terms
Design, Experimentation, Human Factors, Languages, Verification

Keywords
Content Summarization, Literature Search, Natural Language Processing, PubMed, Tag Cloud, Tagging, Text Mining, Visualization

1. INTRODUCTION
While continual improvements in search technology have made it possible to quickly find relevant information on the Web and in other primarily text-based knowledge repositories, the presentation of query results still leaves much to be desired. Few Web applications do anything to organize or to summarize the contents of such responses beyond ranking the items in the list. Thus, users may be forced to scroll through many pages to identify the information they seek and are generally not provided with any way to visualize the totality of the results returned.

Tag clouds are visually-weighted renditions of collections of words (tags) that can be used to represent the concepts present in large collections of information [2]. Tags may be assigned to information resources manually or by automatic indexing. The qualities of associations between each tag and the entity it describes, such as frequency or recency, are visually represented with variable font sizes and colours (e.g. more frequently used tags might be larger or brighter than less frequently used tags). Hyperlinks to the resources described by each tag are often provided for navigation. In recent times, tag clouds have emerged as an important new interface paradigm, quickly gaining popularity in projects such as Flickr [3] and Connotea [1] that need to find visually appealing ways to summarize vast amounts of information.

In this paper, we demonstrate the application of tag clouds to the summarization and navigation of web search results. Using the application we developed, PubCloud, we show that tag clouds enable users to obtain a visual overview of all search results and to use that summary to navigate to relevant subject matter that otherwise would be hidden deep down in the response list.

WWW 2007, May 8–12, 2007, Banff, Alberta, Canada.
ACM 978-1-59593-654-7/07/0005.

2. APPLICATION ON BIOMEDICAL TEXT
The advance of biotechnology has lead to the enormous growth of experimental data, as well as a dramatic increase in the number of scientific publications. PubMed, which is part of the National Center for Biotechnology Information (NCBI), is a centralized repository that indexes millions of biomedical publications [4]. Responses to queries over the abstracts included in PubMed are currently presented in ranked lists that are similar to the responses of most Web search engines. PubCloud is a Java Servlet that queries PubMed for scientific abstracts and summarizes the responses with a tag cloud interface (Figure 1). Implementation details and a usability assessment are summarized below.

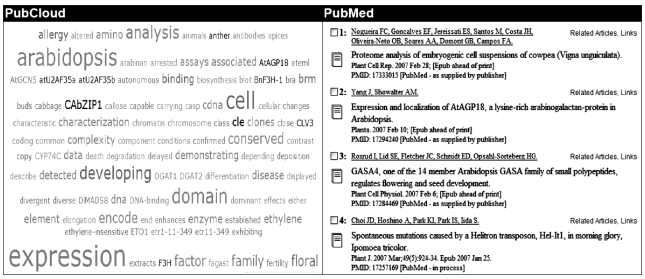

Figure 1. PubCloud offers a summarization of the literature abstracts returned by the PubMed search engine.

2.1 PubCloud Implementation
To generate a tag cloud from the abstracts returned by a PubMed search, two stages of text processing are performed. First, for each abstract, uninformative common words, such as "the", "which", and "it", as well as punctuations and symbols, such as periods, commas, apostrophes, and dashes, are removed. Second, words are parsed from the remaining text and are stemmed, using Porter's Stemming Algorithm [5], to remove suffixes. For example, the word "functional" becomes "funct", and the word "development" becomes "develop". The stemmed words are used as the tags in the tag cloud.

After generating the tag list describing the query response, the relative frequency and recency for each tag is computed.

Frequency is determined by the number of occurrences of each tag and is represented using font size. Recency is based on the average publication date for the abstracts in which a tag appears and is illustrated with different font colours ranging from bright red for the most recent to dark grey for the oldest.

Only tags that are at least 10% as frequent as the most frequent tag are displayed. On each tag, a mouse-over feature displays a list of words that share the same prefixes and a hyperlink links to the set of PubMed abstracts containing it.

3. USABILITY EVALUATION

To assess the usability of PubCloud versus PubMed, we designed a series of questions for users to answer using both interfaces. All questions in the survey were about general topics of plant development. To answer the questions, the users were provided with the output from both PubCloud and PubMed to queries for "terminal flower" and "Arabidopsis". The participants answered the questions in either the abstract-view of PubMed (with all hits in one single page) or the tag cloud view of PubCloud. The survey was run with 20 people of equal number of genders, all of whom were not experts in plant development but had some background in biology. One half of this group answered the questions with PubMed and the other half PubCloud. The number of correct answers was counted for each question, as well as the time spent answering the question.

3.1 Results

The metrics for each question in the survey were summarized in three major categories and are discussed below.

3.1.1 Correctness and quality of the answer

For questions that fall into the descriptive class, for example, "Is TFL a transcription factor?", the quality of the answers was generally higher in PubCloud than in PubMed. Conversely, questions that required the user to identify relationships between multiple concepts, for example, "Name three other genes involved in this process", were better answered using PubMed.

3.1.2 Time spent answering the question

Overall, users spent less time using PubMed to answer questions than using PubCloud. However, similar to the trends observed in the correctness and quality of answers, PubCloud users were able to answer descriptive questions faster than PubMed users. For relational questions, however, PubCloud users spent almost twice as long answering than those who used PubMed.

3.1.3 Degrees of helpfulness, satisfaction and understanding

In addition to the biological questions, users were asked to give their overall impression of the two interfaces. Users ranked PubCloud lower than PubMed regarding the degree of helpfulness. Although they spent more time and found the interface less helpful, the users ranked PubCloud with higher level of satisfaction. In terms of the level of understanding, the users ranked both PubMed and PubCloud equally.

4. DISCUSSION AND CONCLUSION

By summarizing the text content of web pages returned by a query, tag clouds offer not only an overview of the knowledge represented in the entire response, but also an interface that enables users to navigate to potentially relevant information hidden deep down in the response list.

Figure 2 provides a comparison of PubMed and PubCloud along six axes, including answering descriptive questions, answering relational questions, speed, understanding, satisfaction, and helpfulness.

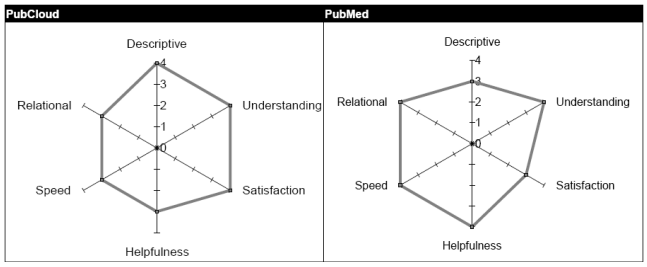

Figure 2. The 'descriptive' and 'relational' axes describe the correctness and the quality of the answers provided by the users. The 'speed' axis was generated based on the amount of time users took to answer each question. The 'understanding', 'satisfaction', and 'helpfulness' axes were based directly on users' responses to questions about the interfaces.

As our results suggested, tag clouds are not a panacea for the summarization of web search results. Although they do provide some improvements in terms of summarizing descriptive information, tag clouds do not help users in identifying relational concepts, and in fact, slow them down when they need to retrieve specific items. Nevertheless, the widespread adoption of tag clouds and the generally positive impression of PubCloud from our users indicate that tag clouds fulfill an empty niche in the current ecosystem of Web interfaces.

Future improvements to PubCloud may include the use of clustering methods to group similar tags, and allowing more flexible visualization controls over font size, colours, hyperlinks, and location on the page. Since the development of tag clouds is ongoing and their application to the summarization of text content is still not well understood, we expect rapid improvement and increased utilization in the future.

5. ACKNOWLEDGMENTS

We thank the volunteers who participated in the evaluation study. BYLK is funded by iCAPTURE Centre for Cardiovascular and Pulmonary Research. TH is supported by CIHR/MSFHR Strategic Training Program in Bioinformatics. BMG is supported by an award from Genome British Columbia. MDW is supported in part by awards from Genome Canada and Genome British Columbia.

6. REFERENCES

1. Connotea: free online reference management for clinicians and scientists, http://www.connotea.org/.
2. Tag cloud - Wikipedia, the free encyclopedia, http://en.wikipedia.org/wiki/Tag_cloud.
3. Welcome to Flickr!, http://www.flickr.com/.
4. McEntyre, J. and Lipman, D. PubMed: bridging the information gap. *Cmaj*, *164* (9). 1317-1319.
5. Porter, M.F. An algorithm for suffix stripping. *Program*, *14* (3). 130-137.