Colored-sketch of Text Information

Beomjin Kim
Indiana University – Purdue University
kimb@ipfw.edu

Philip Johnson
Indiana University – Purdue University
johnpr01@holmes.ipfw.edu

Adam S. Huarng
California State University Los Angeles
ahuarng@calstatela.edu

Abstract

This paper presents an information visualization method, which transforms text into abstracted visual representations. The proposed color-coding algorithm converts text into a sequence of colored icons that inform users about the distributional patterns of given queries, as well as the structural overview of a document simultaneously. By presenting the compact, but instructive visual abstraction of texts concurrently, users can compare multiple documents intuitively while alleviating the need to reference the underlying text. The system provides interactive navigation tools to support users' decision-making processes – including multi-level viewing, a tree hierarchy recording previous search activities, and suggestive words for refinement of the search scope. An experimental study evaluating this visual approach for delivering search results has been conducted on text corpora in comparison with a traditional information retrieval system. By informing search results to clientele in a perceptive form, the users' performance in obtaining desired information has been improved, while maintaining the accuracy.

Keywords: visualization, information retrieval, color-coding, user interface, decision-support system

Introduction

The development of the Internet and digital systems has allowed people unimpeded access to a massive collection of digital information. However, the explosive rate of growth in recent years and the complexity of the information have made it difficult to pinpoint desired information. The conventional search supporting tools, where the users still heavily depend on reading, are not an optimal approach in exploring huge volumes of information. Development of effective solutions in finding relevant information from such a large compilation of data is an important task and is in high demand. In response to the request, researchers have developed methods that present information through visual abstraction. Information visualization has proven to be an effective technique for the navigation of huge information spaces where the users swiftly acquire insight into information with their cognitive activities (Card, Mackinlay, & Shneiderman, 1999). Some document visualization methods have concentrated on presenting sets of outstanding topics of huge corpora through intuitive visual representation (Wise, Thomas, Pennock, Lantrip, Pottier, Schur, & Crow, 1995). Other visualization systems allow users to investigate overall structure and detail information concurrently (Lamping, Rao, & Pirolli, 1995). However, the existing systems are expecting further improvements in delivering distributional behavior of queries and presenting multiple articles on a limited screen space for comparison.

This study has developed a system called Query Fingerprinting (QF) that informs the frequency of queries, distributional information about the queries, and the segmental structures of the document simultaneously. By visualizing a lengthy document into a sequence of icons, this compact, but informative abstraction allows users to compare multiple search results intuitively and swiftly while minimizing the need to reference the underlying contents. Applying this paradigm to other business models - Internet and Digital Library systems that demand more effective informing solutions - will contribute for the improvement of search efficiency in those environments.

In order to evaluate the effectiveness of the developed technique as a search-supporting tool, an experimental study has been conducted to compare the efficiency of the QF system over a conventional summary-based (SB) system. The experimental results showed that the users who employed the QF system showed far better performance, while maintaining the accuracy, in finding relevant information than those who utilized the SB retrieval system. In the post-experimental survey, users show a high degree of satisfaction using the QF system.

Section Editor: Amanda Spink

Related Work

Development of techniques for the presentation of search results in a simple yet instructive configuration has been an important and challenging research area. With the recent rapid growth of information, the effectiveness of the search-supporting tool is of major importance in search efficiency. Numerous approaches have been developed to support users' search activities that can be classified into two groups, summarization and visualization. This paper proposes a method that presents information through visual representations, thus focus is on reviewing previous literature pertaining to visualization.

The rank-based text retrieval system is one of the conventional methodologies used to present retrieved information to users. These systems frequently display a list of titles of retrieved information in order of document relevancy where documents are ranked based on the overall similarity of the document or keywords to a given query set (Cooper, Gey, & Dabney, 1992; Fuhr & Muller, 1987). The rank-based information retrieval (IR) system commonly presents the order of articles with unknown algorithms that do not reflect users' preferences.

One of the most common IR systems shows summaries of relevant documents as a search result. The summary of a document can be constructed by extraction of focused sentences of an article, using language generation or applying artificial intelligence (Tombros & Sanderson, 1998; McKeown & Radev, 1995; Callan, 1994). The summary-based methodology is constrained with respect to the inability to deliver the authors' intention through the short abstract. When the document is lengthy, the summary usually concentrates on the main topic while ignoring a subtopic that could include the users' interests. In addition, both the rank-based and summary-based approaches have a weakness in informing the distribution of the query in the article to the user, which is valuable information in the search process.

Visualization is an effective technique that displays large volumes of information in an intuitive format for easy understanding of the information via perceptual cognition. Various visualization methods have been proposed to provide efficient mechanisms for navigating extensive digital workspaces. Studies have shown that an appropriate visualization technique can increase users' search speeds and accuracies in judging the relevancy of documents (Veerasamy & Heikes, 1997). Such a visualization technique concentrates on showing a portion of the information at a great level of detail while maintaining the overall structure of the information (Robertson & Mackinlay, 1993; Lamping, Rao, & Pirolli, 1995; Furnas, 1986). A recently proposed scrollbar-based visualization utilizes the location of query occurrences by displaying a small icon on top of the scrollbar. Query occurrences are then highlighted throughout the document to assist in finding the location of the query in the text (Byrd, 1999). These visualizations somewhat address limitations in the summary or rank-based approach, but do not have the capability of showing multiple documents simultaneously.

Information clustering is another approach for supporting navigation of a large data collection. Related documents are clustered together whose notable characteristics are visualized using a mixture of attributes (Au, Carey, Sewraz, Guo, & Rüger, 2000; Shneiderman, Feldman, Rose, & Grau, 2000). Allen, Obry, & Littman (1993) developed a visualization method that displays a hierarchical cluster of the relevant documents to a query. The Scatter/Gather algorithm allows users dynamic clustering and refining of ranked search results (Hearst & Pederson, 1996). Also, the Envision system, developed by Nowell, France, Hix, Heath, & Fox (1996), graphically displays various document attributes via a matrix of icons to support users' searches. These techniques have contributed in displaying important features of a cluster compactly and allow users interactive narrowing of their search scopes. Still, the information clustering approach does not show the distributional behavior of queries in a document through its visual representation. For a closer investigation of the document, users still depend on their slower non-perceptual approach: reading. Users will spend a fair amount of time reading a lengthy document where only a portion of the article may show relevance to the search terms.

TileBars shows the frequency of search terms in a document associated with sub topical units, and makes possible the comparison of multiple documents concurrently (Hearst, 1995). This well-known visualization technique for text documents still has room for improvement. Information could be abstracted to a greater extent while providing more information to users, and queries' distributional information should be visualized related to the length of the text segment. Assume that two text segments show the same number of query occurrences but have significantly different lengths. In this case, the query will be more scattered in the lengthy segment than in the shorter segment where queries will be more localized.

Research Goals

The main goal of this research is the development of a method that presents search results more effectively to clientele. In order to address the limitations in conventional approaches, the developed system informs the retrieved results through a visual abstraction. Instead of analyzing the search results, by exploiting the assets of perceptual cognition, the expected search speeds of users will significantly improve while maintaining the level of accuracy in finding the relevant information. In addition, the QF system achieves more data abstraction while providing more information than existing search-supporting tools that use a visual approach in presenting search results. Although evaluation of the relationship between the visual compression and the improvement of search speed has not been explored, this visual compression will be beneficial in navigating huge data collections.

To investigate the effectiveness of the developed method as an informing tool, an experimental study was conducted to compare and contrast this method to a traditional approach. Two hypotheses were generated; stated in the null form, they are as follows:

Hypothesis 1: The total number of articles analyzed per participant using the QF method will not differ from the number analyzed using the SB method.

Hypotheses 2: The accuracy of finding relevant documents using the QF method will not differ from the accuracy of using the SB method.

Methodology

A two-tier process, which consists of information parsing and visualization, converts the text into a visual notation. The information parsing reduces the document's complexity to expedite the follow-up visualization process.

Information Parsing

In order to generate a visual abstraction of the text, while reflecting localized query information, a segmentation process partitions a document into multiple paragraphs. The paragraph demarcations, dictated by the author's structure, are utilized for this segmentation. Segmentation can generate unexpected paragraphs when it meets headings or itemized sentences. These unexpected segments are part of an adjacent paragraph and should be analyzed as part of that paragraph. Besides, the visualization typically maps such unexpected segments to a tiny visual notation, which are hardly perceptible by the users. In order to reduce these artifacts, single sentence segments are attached to the following paragraph when there is one available, if not, that segment is concatenated to the previous paragraph.

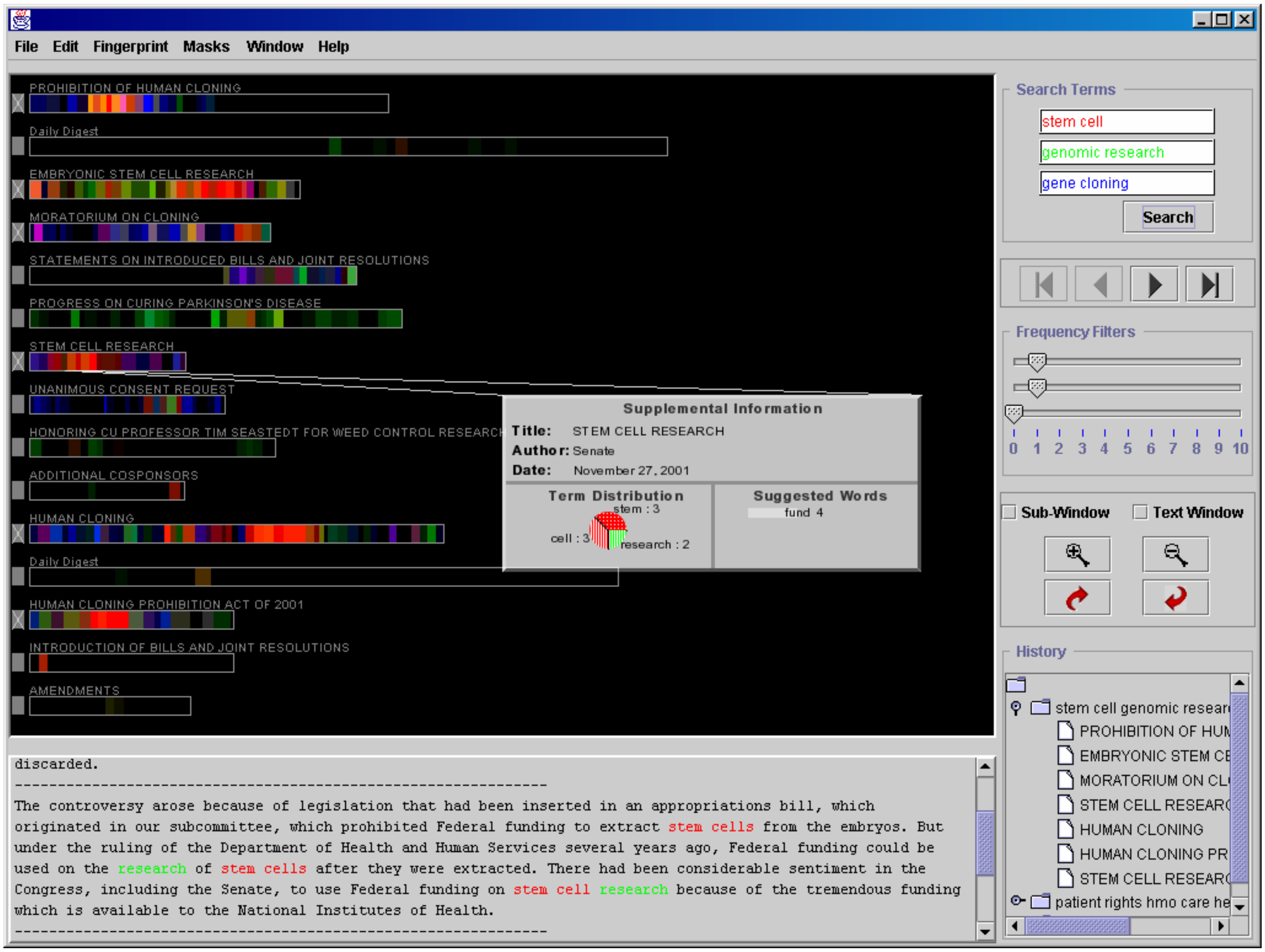
Analyzing the distributional information of queries in a document involves extensive amounts of processing time, which is directly related to the document's complexity. The reduction of this processing time is an important factor in the construction of an efficient IR system. To decrease the complexity of a document, the following procedures are applied to the segmented document. A tokenization procedure breaks each paragraph (P_i) of the document into a set of terms (T_i). The tokens T_i include various stop-words that are not used to facilitate the content of the document's connotation, but are used to establish the context in which the author desires to disseminate his/her information (Paice, 1994). The stop-word removal process eliminates the stop-words from T_i to decrease the document's complexity without affecting the content of the document. Finally, a suffix stemming procedure conflates a group of related words into a single stem word. To consider the system performance, Porter's algorithm, which removes suffixes without the use of a stem dictionary, was utilized (Porter, 1980). These filtering processes are essential in improving the performance of the following visualization procedure, which requires a substantial amount of computing time.

Visualization

The QF methodology utilizes three attributes: color, intensity, and size, to represent distributional behavior of both the query and document concurrently. To generate color codes related to search terms, a primary color (red, green, or blue) is assigned to a subset of the query. For each paragraph P_i, the frequencies of the search terms are analyzed. The search terms which make a query subset are processed as an OR relationship, so that the frequency of each term in a paragraph is added together to form the frequency of the query subset in a paragraph denoted by $(F^C)_i$.

$(F^C)_i = (\sum_{j=1}^{n} q_j^C)_i$ where q_j^C is the frequency of a query in a paragraph i, C is a query subset {R, G, B}.

Meanwhile, query subsets are processed as an AND relationship by passing through a color mixing procedure. The ratio $(R^C)_i = (F^C)_i / F_i$, where F_i is the number of words in P_i, is mapped to an intensity scale of sixteen different intensities to determine the intensity of each query subset. The higher ratio, $(R^C)_i$, maps to the brighter intensity of the corresponding color. When the $(R^C)_i$ is greater than a predefined threshold value, then the $(R^C)_i$ maps to the highest intensity that is the brightest hardware intensity of the color. This mapping function is repeatedly applied to each query subset for every P_i to produce the three primary color intensities, $I^R{}_i$, $I^G{}_i$, $I^B{}_i$, where i is the i^{th} paragraph in the document, and ^{R}, ^{G}, and ^{B} represent the primary colors, red, green, and blue. By mixing $I^R{}_i$, $I^G{}_i$, and $I^B{}_i$ together, a color Ω_i will represent the query distribution in P_i, that will



stem cell

genomic research

gene cloning

Title: STEM CELL RESEARCH
Author: Senate
Date: November 27, 2001

□ Sub-Window □ Text Window

```
discarded.
------------------------------------------------------------
The controversy arose because of legislation that had been inserted in an appropriations bill, which
originated in our subcommittee, which prohibited Federal funding to extract stem cells from the embryos. But
under the ruling of the Department of Health and Human Services several years ago, Federal funding could be
used on the research of stem cells after they were extracted. There had been considerable sentiment in the
Congress, including the Senate, to use Federal funding on stem cell research because of the tremendous funding
which is available to the National Institutes of Health.
------------------------------------------------------------
```

Figure 1: Example view of the query fingerprinting system.

be painted onto an icon whose width is determined by $f(F_i) = F_i / \lambda$ where λ is a scaling factor between number of words and pixels. This visualization process is iterated for every P_i to create an iconic view of P_i, which is concatenated together to form the QF of a document.

Through the frequency-to-intensity mapping, a paragraph that has higher concentration of query occurrences will be projected to a brighter color. A paragraph that discusses more extensively the user's interest, that is, higher query occurrence in a shorter paragraph, will be displayed brighter on a black background window than a longer paragraph having the same query occurrence would. At the same time, each strip also delivers the query's distributive tendency within a paragraph. For example, when a paragraph contains search terms from only one of the query subsets, the strip will show a primary color depending on the corresponding subset's associated color. Meanwhile, when a paragraph includes occurrences of the subsets associated to the colors red and green, the strip appears yellow with varying intensity depending on the concentration of those two subsets.

The QF system allows users to compare paragraph lengths and document lengths concurrently by displaying the QF of documents top to bottom. Figure 1 illustrates an example view of the QF system that shows multiple documents simultaneously. The first QF of the figure shows much brighter overall intensity at the beginning of the document. Meanwhile, the second QF shows relatively low intensity, which suggests to the user that the corresponding document might not closely relate to his/her interest. The users will start further investigation from the first document, especially from the left half of the document, which shows much brighter intensity than the right side of the document does. These abstracted visual representations will make the users' search processes more efficient while avoiding unrelated documents and also unrelated portions of a document.

Supporting Systems

To assist users' search activities, the QF system provides several supporting functionalities. The multi-level viewing allows users to examine retrieved results from different viewpoints. The initial QF displays a general outline of the document. When a user needs further investigation of a specific segment before making a decision, by pressing the left mouse button over an icon, a pop-up window shows supplementary information of the document and the paragraph associated to the icon. The sub-window provides title, author, and date of publication, along with detailed distributional information about the search terms within the paragraph (Figure 1). It also provides other significantly related terms in the paragraph that will help users to understand the content of the paragraph better and refine their search scope (Spink, Jansen, Wolfram, & Saracevic, 2002). When the users press the right mouse button, the corresponding paragraph is displayed at the bottom of the window where given queries are highlighted with the corresponding query's color to assist users in rapidly locating queries in the document. To support the users' understanding of topical transitions between paragraphs, the selected paragraph and any occurring adjacent paragraphs are presented.

Selection of the proper search terms is important to get better search results. Users will commonly experience very low hits and large hits in their search processes. When there are an excessive number of search results, the limited screen space does not allow users to compare multiple documents at once. The QF offers a feature that makes it easy to organize candidate search results. When the user selects a document that needs further investigation, the selected document is attached to a search result tree that organizes the selected documents associated with given queries. The lower left corner of Figure 1 shows an example of the tree-shape structure of search activities tied with search terms. This feature helps users revisit different types of documents later, even though the search scope is refined during the decision making process.

Experiments

In order to evaluate the effectiveness of the QF method as a search-supporting tool, we have conducted experimental studies on text corpora. To assess the hypotheses defined in the previous section: Research Goals, the swiftness and correctness in finding relevant articles among the text corpora were compared. The two applications used were the QF methodology and the traditional SB approach, respectively.

Experimental Environments

The experiment was conducted in an isolated PC-Lab where each computer is equipped with an 866 MHz Intel CPU and 256MB main memory. Both applications have been developed with J2EE^{TM} Platform. The QF application has three major components, which are a main window displaying QFs, a sub-window showing a portion of the text document, and a control window that includes various filtering and decision supporting tools. In this experiment, for accurate evaluation of the QF method, participants have analyzed a data set solely based on the appearance of QFs without using the supporting functionalities in the QF system. The SB system is developed based on the best-matching sentence extraction approach. In the SB application, query frequency in each document is analyzed and the three sentences that include the highest frequencies of the query are selected as a summary of the document. When a user clicks the mouse on a summary of article, the contents of the document are displayed at the bottom of the GUI from the beginning of the document. In both systems, terms contained in a query subset in an article are displayed with the corresponding query subset's color to help users find the query's locations in a document.

Data Sets and Participants

Experimental studies were conducted on a set of articles that were collected from the Library of Congress database, *Congressional Record* (*CR*), which contains an account of everything that is said and done on the floors of the House and Senate (Library of Congress Congressional Record). The participants' prior knowledge of the corpora's subject matter may be a factor influencing their decision-making speed. To acknowledge diversities of interest, nine different data sets, with mutually exclusive topics, were prepared for the experiments. Topics for the data sets were selected from social issues, heavily discussed in the mass media at the time of data collection (Table 1). The search terms were defined by the most frequently occurring words in related articles, which were available to the public as an electronic or hardcopy. A set of queries were supplied to the 2001 Congressional Text Archive from which the first two hundred texts were returned grouped by date with the most current texts first. From the set of two hundred texts the first fifty texts, which after document parsing (stop-words removal and suffix stemming) contained between 800 and 5000 words were collected for the data set.

Data Sets	Queries	Number of highly-related documents	Number of related documents	Number of non-related documents
1	(Patient, Rights) (HMO, Care) (Health, Insurance)	16	9	25
2	(Tax, Cuts) (State, Income) (Federal, Estate)	14	20	16
3	(Immigration, Visa) (Student, Work) (Temporary, Migrant)	11	20	19
4	(Free, Trade) (Favorite, Nation) (Agreement, Status)	9	13	28
5	(Stem, Cell) (Genomic, Research) (Gene, Cloning)	9	4	37
6	(North, Atlantic) (Treaty, Organization) (United, Nations)	7	7	36
7	(Economic, Stimulus) (Federal, Package) (Interest, Rate)	4	27	19
8	(World, Bank) (Debt, Cancellation) (International, Monetary, Fund)	4	19	27
9	(Kyoto, Protocol) (Global, Environment) (Climate, Control)	4	19	27

Table 1: Experimental data sets.

A query set and a group of related articles were presented to two independent readers on each day for the relevancy judgment. Both independent reviewers read every article and assigned a relevancy score for each one. Based on the reviewer's judgment, five points were assigned to articles most relevant to the given query set, three points for medium relevancy, and one point to the documents having least relevancy. The scores on an article from the two readers were averaged to represent the Experimental Standard Score of the corresponding document related to the given query.

Twenty-one experimental participants have been recruited from the first author's university. They were all Computer Science majors, native English speakers, had no difficulties using the computers, no problems in color perception, and have experience using some type of search applications.

Experimental Tasks and Procedure

A randomly selected query was given to participants who were asked to identify as many relevant documents to the query set as possible in a ten-minute period. After investigating each document, these participants assigned a relevancy score of the document to the query set using the same scoring scale that was mentioned above. If a participant could not finish reviewing a document within ten minutes that document was not counted for further consideration. For the fairness of comparison, each participant was given the exact query set supplied to the *CR* database to construct the experimental data sets. The number of documents in the *CR* database is so vast that the time needed for retrieving these documents over the network can be a variance factor in this experiment. For a more accurate comparison between the two systems, the fifty documents were retrieved from the database and were locally saved, so that the network traffic was not the variance in this experiment.

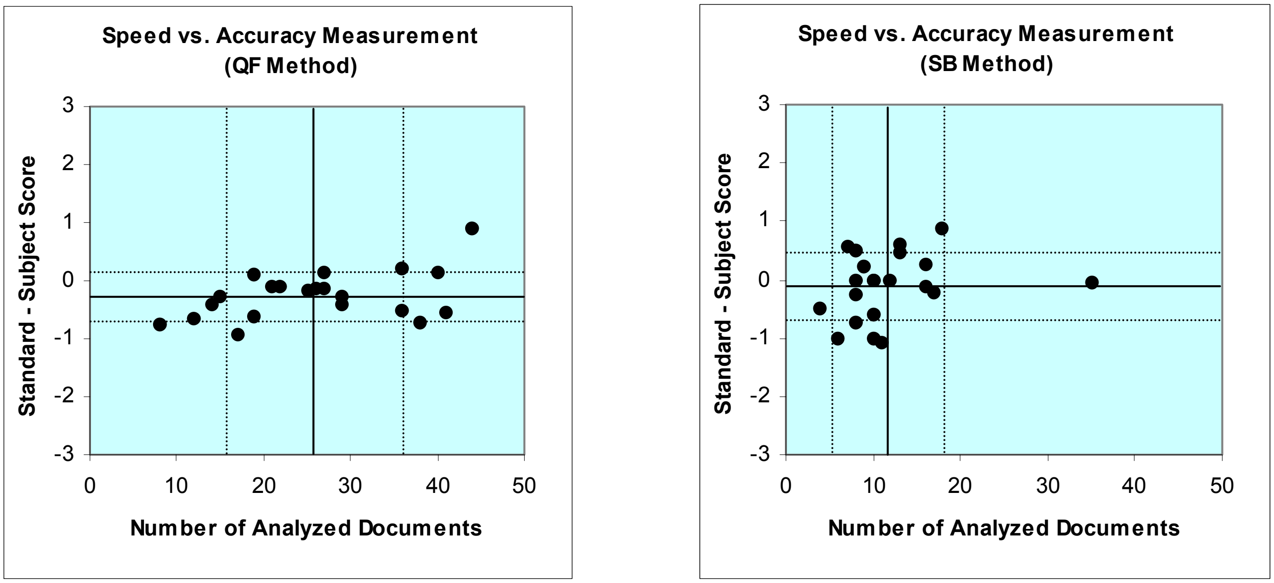


Figure 2: Experimental results for evaluating the QF system.

Although both systems were uncomplicated and intuitive to use, an orientation to using both systems was given to each participant before starting the experiment. This orientation explained the use of the content viewing functionalities that are provided by both systems and the meaning of color-codes and the length of strips in the QF system. Using two different experimental data sets, each participant found the related documents with the SB system for a data set and the QF system for another.

The accuracy at which a participant could judge related documents was measured by computing the score difference between the Experimental Standard Score and the score assigned by experiment participants. The speed of exploring the data collection was evaluated by measuring the number of articles analyzed within the experimental time period.

Experimental Results

The following shows the experimental results comparing multiple users' speed and accuracy in judging the relevancy of documents in both the QF method and the SB method. The chief variance factor in this experiment was the participant's swiftness at reading an article to judge its relevancy. For instance, some participants felt that they had to read significantly more portions of a document than others did before they could make a decision about the relevancy of the document.

A t-test was performed to measure the difference between QF and SB methods in terms of the number of articles that have been analyzed. The results showed that the means were significantly different (p <= 0.000) (for all hypothesis tests, we assumed that $\alpha = 0.05$). The participants, who used the QF method, analyzed more articles than those who used the SB method. The result supports the rejection of null Hypothesis 1.

Hypothesis 2 stated that the accuracy of finding relevant documents using the QF method would not differ from the accuracy of using the SB method. The results of the t-test showed that there were no differences between the two methods ($p < 0.331$). Consequently, Hypothesis 2 cannot be rejected. Table 2 shows the results of the t-test. Figure 2 shows the speed and accuracy comparisons between the two methods where the x-axis represents the document analyzing speed and the y-axis shows the accuracy in judging the relevancy of the article. The solid lines in the figures represent the mean values and the dashed lines represent the standard deviation for accuracy and swiftness in judging the documents.

Post Experimental Results

The effectiveness of QF system as a search-supporting tool was also demonstrated by a post experimental survey. The questionnaire asked the participants' opinions about using the QF system and comparison with the SB system. Although, 72% of participants stated that at first they were more comfortable using the SB system whereas only 11% of the participants felt more comfortable using the QF system, by the end of the experiment, 77% of the participants stated that the QF system was more valuable than the SB system in their search process and 66% of the participants would prefer to use the QF system in the future.

For the questions regarding the ability to understand the denotation of a color code, 61% of the participants stated that they encountered no major difficulty envisioning the search terms' correlation to the color codes, whereas 22% stated that they had difficulty understanding the meaning of the blended colors.

The participants also replied positively to the features in the QF system regarding visualizing the structure of a document in a simple and understandable manner. When participants were questioned about the QF system's ability to assist in the recognition of document structure without referencing the actual contents 72% of the participants agreed that the QF was supportive. Also, 88% of the participants said that they could easily compare multiple documents at a glance.

Finally, when asked their opinion about the ability of the QF method to assist users' search and decision-making processes when applied to an Internet search engine or Digital Library systems 89% of the participants agreed that the QF method would be helpful, 11% of the participants had a neutral stance, and none of participants felt that the QF would not be helpful.

Discussion and Future Research

In this experimental study, the participants who used the QF method in judging the relevancy of articles showed significant improvement in their document analyzing speed. The developed algorithm transforms expansive information into compact visual abstractions that allows understanding of the distributional behavior of search terms and structural layout of articles. This compact, but intuitive visual representation makes it possible to review a large number of documents swiftly through users' perceptual cognitions, while reducing the reference to actual information. In addition, the color-coded iconic representation of text segments assists users in promptly accessing a location within the article where the users' interests lie. Development of this technique allows users to compare multiple search results at the same time on a limited space. This results in the improvement of search speed and is an essential mechanism in exploring current information workspaces that are expanded exponentially.

Meanwhile, the accuracy of the relevancy scores shows a slight overestimation. The participants who used the QF system showed a tendency to classify an article as more relevant than those who used the SB system. One of the reasons for the overestimation could be the way in which the users judge the relevancy of the documents. The participants using the QF system commonly started their document investigation from a portion of the document where a high query concentration occurs. Based on that specific portion of the document the participant may assign a relevancy score that, taking into account the entire document, possibly overestimates the document's actual relevancy.

Users' knowledge, related to the topics, can be another factor affecting their evaluation speed and accuracy. The users could find it easier to judge the relevancy of an article if the article, in some way, contained information that was of interest to them. For a more non-biased evaluation of the QF system as a search-supporting tool more usability testing needs to be done with different experimental environments. For example, allow the user to choose which query they would like to use based on their familiarity to, and interest in, the topics. Also, performance comparisons to similar methods that represent documents through visualization are another way of evaluating the

	Mean		Std. Deviation		Pooled Variance Estimates		
Measure	QF	SB	QF	SB	T Value	D.F.	Significant Level
Speed*	25.95	11.86	10.28	6.48	5.315	40	0.000

* Number of analyzed articles

	Mean		Std. Deviation		Pooled Variance Estimates		
Measure	QF	SB	QF	SB	T Value	D.F.	Significant Level
Accuracy *	-0.25	-0.1	0.42	0.56	-0.983	40	0.331

* Experimental Standard Score – Subject Score

Table 2: T-test comparing differences in speed and accuracy.

validity of the QF method. The number of articles within an experimental data set was not large enough to investigate the effectiveness of the QF system in exploring a very large information workspace. Currently, we are expanding this study to a TREC-5 test collection where the performance of the QF system will be compared to other approaches that present search results through visualization. Finally, an investigation of the effect of using other search supporting features in the QF system can also be explored and evaluated.

Although most of the text information includes predefined paragraph boundaries, some text documents do not contain such a structure. Especially, huge and heterogeneous information on the Internet commonly have itemized or continuous blocks of text information. To visualize unformatted information, while reflecting the localized queries' distributional information, the QF method should be equipped with an automated segmentation capability. Adopting existing text segmentation algorithms or creating pseudo-paragraphs by regrouping sentences in the article could solve the problem.

The QF method can be applied in visualizing clusters of information. Related documents are grouped together to form a cluster, and the QF of each article is compressed more intensely. By placing the condensed QFs adjacently based on their relevancy to search terms, a group of QFs will show the distributional pattern of the query for individual documents and a cluster of articles simultaneously. This visual representation will address the drawback of current visualization methods in displaying clusters of information.

While this study has shown the helpfulness of the QF system as a search-supporting tool, more research should focus on the following areas. Although users can provide multiple search terms for each query subset, and applying more than three query subsets are not common search behavior, the number of primary colors could be a constraint in using the QF system. Introduction of the other attributes into the QF system could be a possible solution to tackle this restriction. The incorporation of a document ranking and multi-level zooming functionality to the QF system will be another beneficial utility in users' search processes. The level of intensity and color distribution can be valuable properties to rank the retrieved information with respect to the search terms.

A recent study regarding search behavior on the Internet suggests that over one half of users do not look over retrieved results beyond the first page (Jansen, Spink, & Saracevic, 2000). One effective approach in accordance to this search activity will be the development of methods that classify the most relevant articles to the users' interests from the collection of related information. Introduction of zooming capabilities to the OF's visual abstraction will be another methodology to meet users' search patterns. The multi-level zooming functionality makes it possible to compare, simultaneously, numerous articles from different levels of abstractions and to visualize the overall layout of very lengthy information within a limited screen space.

Finally, applying the QF methodology to the other emerging information workspaces such as an Internet search or Digital Library systems will be another challenging task. The abstracted visual representation from the QF methodology will address the limitations in existing search engines that display a list of titles and summaries as a search result.

Conclusion

This paper has introduced a visualization method that displays text information compactly, but informatively through a color-coding algorithm. This visual abstraction makes it possible to show detailed information – including queries' frequencies, their correlations, and the lengths of paragraphs and documents – while maintaining the overall structure of the information in the same view. By presenting distributional and structural information with respect to a segment of the article, the QF method addresses the deficiencies of the existing text visualization methods. The efficiency of the QF technique in identifying desired information from a large data collection has been acknowledged in our experimental study. In summary, the QF can be an effective informing tool in the search process that will eventually support users' decision-making processes, and save their time and effort by alleviating the need to reference any of the irrelevant information. Applying these techniques to emerging on-line environments will be beneficial to the future IR system.

Acknowledgments

We would like to thank all students who participated the experimental study for their help and suggestions. The first author's research was sponsored in part by a PRF Summer Faculty Grant from the Purdue Research Foundation.

References

Allen, R., Obry, P., & Littman, M. (1993). An interface for navigating clustered document sets returned by queries. *Proceedings of the Conference on Organizational Computing Systems*, 166-171.

Au, P., Carey, M., Sewraz, S., Guo, Y., & Rüger, S. (2000). New paradigms in information visualization. *Proceedings of the Twenty third Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 307-309.

Byrd, D. (1999). A scrollbar-based visualization for document navigation. *Proceedings of the Fourth ACM Conference on Digital Libraries*, 122-129.

Callan, J.P. (1994). Passage-level evidence in document retrieval. *Proceedings of the Seventeenth Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 302-310.

Card, S.K., Mackinlay, J.D., & Shneiderman, B. (1999). Readings in information visualization using vision to think. *Morgan Kaufmann*, 1-34.

Cooper, W.S., Gey, F.C., & Dabney, D.P. (1992). Probabilistic retrieval based on staged logistic regression. *Proceedings of the Fifteenth Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 198-210.

Fuhr, N. & Muller, P. (1987). Probabilistic search term weighting – some negative results. *Proceedings of the Tenth Annual ACM SIGIR Conference on Research and Development in Information Retrieval*, 13-18.

Furnas, G.W. (1986). Generalized fisheye views. *Proceedings of the ACM Conference on Human Factors in Computing Systems*, 16-23.

Hearst, M.A. (1995). Tilebars: Visualization of term distribution information in full text information access. *Proceedings of the ACM Conference on Human Factors in Computing Systems*, 59-66.

Hearst, M.A. & Pedersen, J.O. (1996). Reexamining the cluster hypothesis: scatter/gather on retrieval results. *Proceedings of the Nineteenth Annual ACM SIGIR Conference on Research and Development in Information Retrieval*, 76-84.

Jansen, B. J., Spink, A., & Saracevic, T. (2000). Real life, real users, and real needs: A study and analysis of user queries on the Web. *Information Processing and Management*, 36(2), 207-227.

Lamping, J., Rao, R., & Pirolli, P. (1995). A focus+context technique based on hyperbolic geometry for visualizing large hierarchies. *Proceedings of the ACM Conference on Human Factors in Computing Systems*, 401-408.

Library of Congress Congressional Record. (2001). *http://thomas.loc.gov/home/abt.cong.rec.htm*.

McKeown, K. & Radev, D.R. (1995). Generating summaries of multiple news articles. *Proceedings of the Eighteenth Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 74-82.

Nowell, L.T., France, R.K., Hix, D., Heath, L.S., & Fox, E.A. (1996). Visualizing search results: some alternatives to query-document similarity. *Proceedings of the Nineteenth Annual ACM SIGIR Conference on Research and Development in Information Retrieval*, 67-75.

Paice C.D. (1994). An evaluation method for stemming algorithms. *Proceedings of the Seventeenth Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 42-50.

Porter, M.F. (1980). An algorithm for suffix stripping. *Program*, 14(33), 130-137.

Robertson, G.G. & Mackinlay, J.D. (1993). The Document Lens. *Proceedings of the Sixth Annual ACM Symposium on User Interface Software and Technology*, 101-108.

Shneiderman, B., Feldman, D., Rose, A., & Grau, X.F. (2000). Visualizing digital library search results with categorical and hierarchical axes. *Proceedings of the Fifth ACM Conference on Digital Libraries,* 57-66.

Spink, A., Jansen, B. J., Wolfram, D., & Saracevic, T. (2002). From E-sex to E-commerce: Web Search Changes. *IEEE Computer*, 35(3), 107-111.

Tombros, A. & Sanderson, M. (1998). Advantages of query biased summaries in information retrieval. *Proceedings of the Twenty First Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2-10.

Veerasamy, A. & Heikes, R. (1997). Effectiveness of a graphical display of retrieval results. *Proceedings of the Twentieth Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 236-245.

Wise, J.A., Thomas, J.J., Pennock, K., Lantrip, D., Pottier, M., Schur, A., & Crow, V. (1995). Visualizing the non-visual: Spatial analysis and interaction with information from text documents. *Proceedings on Information Visualization*, 51-58.

Biographies

BEOMJIN KIM, Ph.D., is an assistant professor of Computer Science at the Indiana University-Purdue University in Fort Wayne, Indiana, USA. He obtained his Ph.D. in Computer Science from the Illinois Institute of Technology in 1998. His current research interests include information retrieval and visualization, medical image, and computer graphics.

PHILIP JOHNSON is a student member of the Information Visualization Laboratory at Indiana University Purdue University Fort Wayne. Philip is the acting President of the IPFW student chapter of the ACM, and he is the IPFW Computer Science Department representative to the Fort Wayne chapter of Infraguard. His research interests are: Scientific Visualization, Computer Graphics, and Information Security.

ADAM S. HUARNG, Ph.D., is currently an Associate professor at California State University in Los Angeles. His articles have appeared in many information systems journals such as Journal of Computer Information Systems, Information & Management, Information Systems Management and Strategic E-commerce. He can be reached at ahuarng@calstatela.edu.