# Creating Visualisations for Digital Document Indexing

Jennifer Pearson, George Buchanan, and Harold Thimbleby

FIT Lab, Swansea University
{j.pearson,g.r.buchanan,h.w.thimbleby}@swan.ac.uk

**Abstract.** Indexes are a well established method of locating information in printed literature just as FIND is a popular technique when searching in digital documents. However, document reader software has seldom adopted the concept of an index in a systematic manner. This paper describes an implemented system that not only facilitates user created digital indexes but also uses colour and size as key factors in their visual presentation. We report a pilot study that was conducted to test the validity of each visualisation and analyses the results of both the quantitative analysis and subjective user reviews.

**Keywords:** Document Triage, Indexing, Information Visualisation.

## 1 Introduction

Searching for relevant information within a document is a tiresome but necessary procedure in the information retrieval process. In physical books the process of locating relevant information is supported by the index; a classic structure that references key terms within a document for easy navigation. The problems associated with this method are that they are author created which restricts them to static terms that exist when the book is created and that they take time to physically navigate to.

One dynamic user-prompted method of information seeking on electronic documents is text search (Ctrl+f). By facilitating user-defined search terms, Ctrl+f has established itself as being a considerable advantage in electronic documents. However, in many document readers and web browsers this function does not return a list. Instead, it simply steps linearly through the document, highlighting each occurrence one at a time. This sequential interaction is slow and cumbersome if the main aim of the search is to get an overview of the location of text matches in a document. This concept of keyword overview can be a useful feature when performing document triage as it gives the user a solid indication of the areas of the document to observe first.

The limitations of linear search are underlined by a known discrepancy between actual behaviour and self reported use [3]. In fact, these shortcomings often force users to make use of other tools such as Internet search engines to overcome these problems. More sophisticated document reader software such as Preview for the Mac provide search facilities with a different presentation, giving

M. Agosti et al. (Eds.): ECDL 2009, LNCS 5714, pp. 87–93, 2009.

a list of results as well as a small snippet of text surrounding them. Several occurrences of a keyword on the same page then yeild multiple result lines. However, even this improved interaction is not a perfect solution. For example, this type of visualisation relies on the users' ability to mentally 'group' these same-page occurrences to decide where there are relevant clusters of information.

This paper introduces a system that improves information seeking by incorporating both user-prompted searching and indexing in a custom index-builder. The Visual Index System (VIS) both creates user defined indexes and visualises them in different ways to illustrate the relevance of each page/cluster.

## 2 Related Work

Although there has been much research into general document retrieval, there seems to be a relatively limited amount of exploration into within-document search techniques. Harper et al. [2] have investigated the differences between two such systems: FindSkim which has been designed to mimic standard search (Ctrl+f) tools found in most text processing and web browsing applications and SmartSkim, a visual interpretation based on 'relevance profiling'. Their results concluded that SmartSkim; a bar chart visualisation that constructs 'bars' based on word (not phrase) occurrences in text chunks was, in general a more effective method of precision and recall than FindSkim.

Byrd [1] discusses visualisations for whole- and within- document retrieval, making use of multi-coloured 'dots' that appear in the scroll-bar, and that correspond to the locations of keywords within the document. The user can use this display to identify clusters of a keyword by close proximity of dots of one colour.

## 3 Design and Implementation

The main aim of VIS is to allow users to create their own index which in turn gives a visual overview of the most relevant parts of the document based on the keywords entered. It groups the results that appear on the same page (and also occurrences that appear in clustered pages) and visualises their occurrences by means of colour and size. Creating a graphical representation of the data increases cognitive activity allowing users to better understand the underlying

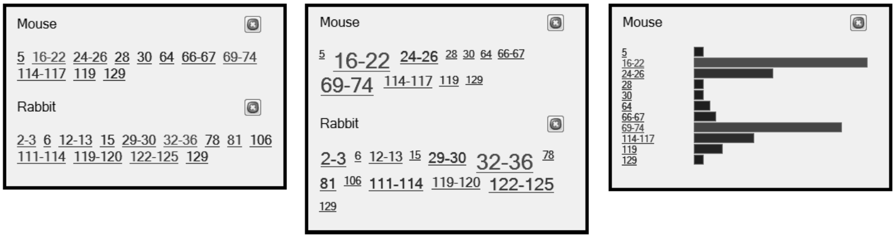


**Fig. 1.** The Three Types of Visualisation: Colour Tag, Tag Cloud and Graph

information [4] giving a clear overview of relevant sections by illustrating where the highest occurrences of each keyword appear.

VIS provides three separate index visualisations (see Fig 1) and allows users to toggle between them easily by means of a radio button set on the taskbar:

**Colour Tag:** The Colour Tag system is the same as a traditional index list layout in terms of size, but we have also coloured each link depending upon the number of occurrences of the word on that page/cluster.

**Tag Cloud:** The Tag Clouds system is an alphabetically sorted, size-weighted list of links that allow users to easily see each page/clusters relevance by means of their size and/or colour.

**Graph:** Harper et al produced the SmartSkim interface (see Section 2) which produced a vertical interactive bar graph representing the document and each section's relative retrieval status values. Working with this idea in mind, we decided to incorporate a simple graph type in the visual indexing solution. The simple horizontal bar chart implemented into the system represents the page/clusters versus the number off occurrences of the keyword/phrase.

The basic approach of the VIS system includes three functional features which are used in all of the visualisations we are exploring:

**Hyperlinks:** Each of the search results in the system will be in the form of a hyperlink which when clicked will take the user to the appropriate page and highlight all occurrences of the keyword/phrase on that page.

**Page Clusters:** One feature that the program possesses is its ability to 'cluster' page hits. For example, in a catalogue when you are looking for all references to sofas the index may look something like this: Sofa: 345, 467, **1067-1098**.

**Tool Tips:** Each link has a tool tip containing the occurrence information of the particular keyword/phrase,which pops up when the mouse hovers over it.

These visualisation methods use two distinct visual cues:

**Colour:** To minimise the time taken for a user to process the information presented by VIS, we used a system of colour coding to indicate areas of the document with most occurrences of the particular keyword or phrase. The colours used for this visualisation have been selected to mimic a temperature gauge; i.e. links with a low number of occurrences will be blue in colour whereas links with a high number of occurrences will be red.

**Size:** In our three new visualisations, we use different size-related cues to indicate the number of matches in a set of pages. In the Tag Cloud mode, the size of the text indicates the number of matches; in Graph mode, the length of the bars; the colour tag mode does not use size.

## 4 Pilot Study

To investigate the issues described in section 3, a pilot study was performed which focused on qualitative data in the form of subjective user questionnaires, as well as quantitative data obtained from task timings. Our hypothesis was that a custom indexing solution would in itself be an improvement over traditional linear indexing and furthermore, that a visual representation (colour and size) would prove to be a valuable asset in triage efficiency. In order to test the effectiveness of the implemented visualisations, two additional searching methods were written as a basis to test against. These two techniques were based on long-established methods of document searching:

**Linear Search:** The linear search method implemented in the program is based on the standard Windows within-document find feature that allows users to sequentially progress through a document one keyword at a time.

**Traditional Indexing:** The traditional indexing method has been designed to look like the classic index structure i.e. all entries being the same size and colour. It does however have the same 'build' feature as the visual index methods and also includes hyperlinks and tool tips.

The participants chosen for the study were selected from a set of postgraduate researchers due to their increased knowledge of online and document searching. In total 14 participants were selected; 11 male and 3 female, all between the ages of 22 and 37 and all with normal colour perception. All recorded data was kept anonymous, the study itself lasted on average around 30 minutes and the participants were given a £5 gift voucher in return for their time.

The structure of the study comprised of a pre-study questionnaire designed to gain general information about each participants searching habits, followed by a set of tasks and finally a post-study questionnaire devised to obtain subjective views with regards to the tested systems. The tasks provided were designed to time how long it takes to find the most relevant parts of a document based on a particular keyword(s). Users were given 3 PDFs and asked to perform 5 separate searches on each; one for each of the search methods. They were then asked to discover the part of a specific document that best matched a given query.

Due to the nature of the study, there were several factors that could affect the time taken to complete the tasks. These include the length of the document and the number of occurrences of the keyword/phrase in the document. To minimise the effects of these factors, each search term was assigned on a latin-square design to balance orderings and the logged time data across the 3 separate PDFs was averaged. To reduce any bias resulting from the different PDFs and any possible learning effects, the orderings of the documents and tasks were varied.

**Table 1.** The Averaged Timed Data (in seconds)

| | Linear Search | Traditional Index | Colour Tag | Tag Cloud | Graph |
|---|---|---|---|---|---|
| AVERAGE PDF 1 | 99.9 | 37.7 | 21.2 | 14.4 | 14.5 |
| AVERAGE PDF 2 | 92.1 | 33.9 | 13.2 | 10.9 | 9.9 |
| AVERAGE PDF 3 | 134.1 | 32.8 | 15.6 | 11.6 | 11.2 |
| AVERAGE ALL | 108.7 | 34.8 | 16.7 | 12.3 | 11.9 |
| SD ALL | 59.33 | 22.78 | 11.18 | 9.54 | 10.20 |

## 4.1 Results

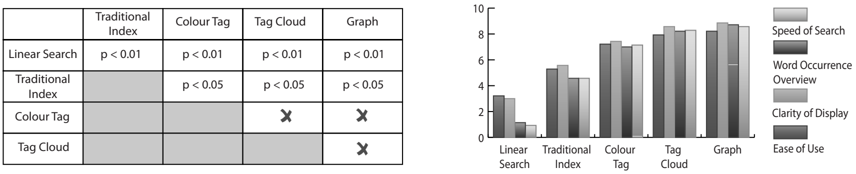
The results of the timed tasks (see Table 1) performed on each search method concluded that the traditional index is approximately 3 times faster than the standard linear search for locating relevant information in a document. Furthermore, it also confirmed that the use of colour and size is a further improvement with the average time for completion being at least half when using the visual systems (Colour Tag, Tag Cloud and Graph) over the traditional index. These statistics are also backed up by the subjective user ratings as shown in Fig 2b.

To assess the statistical significance of the timed test results a single-factor ANOVA test was performed. This test produced results of $p < 0.0001$, $df = \{4, 205\}$, $F = 81.915$ and $F\mathit{crit} = 2.416$, concluding that the resulting data was statistically significant. Due to the non-similar variances of the linear search versus the index builders, a Welch's t-test with bonferroni corrections was conducted upon the data to pin-point the areas of major significance.

The results of the bonferroni t-test (Fig 2a) confirm the significance of the difference between linear search compared against the 4 indexing methods, consistently yielding $p < 0.01$ and t values ranging from 11.470 to 15.030. Furthermore, comparing the traditional indexing method to each of the visual indexing methods (Colour Tag, Tag Cloud and Graph) has also been proven to have statistical significance with $p < 0.05$ and t values ranging from 2.816 to 3.560.

The tests performed clearly defined the major differences between the five implemented systems. Unsurprisingly then, the systems that performed similarly in the time tests (i.e. the visual systems) have yielded non-significant results from this data. A more precise study of these methods will be required to make a more concrete analysis of the differences between the visual index systems themselves.

As well as determining the different speeds in which each system can locate relevant information in a document, the study also determined the precision. In the linear tasks users were able to locate the most relevant section(s) of a document only 40.47% of the time, whereas the traditional index method yielded a result of 73.81%. We applied the Chi-squared test to these two modes, giving a significant result of p=0.002 (chi=9.528). Furthermore, The Colour Tag system allowed users to find the most relevant section 95.23% of the time and the Tag Cloud and Graph methods both resulted in 100% accuracy. It is clear from these results that not only is a custom index builder a large improvement in relevance accuracy than traditional linear searching, but also substantiates the theory that



| | Traditional Index | Colour Tag | Tag Cloud | Graph |
|---|---|---|---|---|
| Linear Search | p < 0.01 | p < 0.01 | p < 0.01 | p < 0.01 |
| Traditional Index | | p < 0.05 | p < 0.05 | p < 0.05 |
| Colour Tag | | | ✗ | ✗ |
| Tag Cloud | | | | ✗ |

**(a)** Statistical Significance

**(b)** User Ratings

**Fig. 2.** Results: a) Performed using student t-test with bonferroni corrections. **x** indicates a non-significant result b) Graph of User Ratings (out of 10)

colour and size add further precision to the search. A global Chi-squared test (all 5 interfaces) produced chi=98.96; $p < 0.00001$.

The averaged user scores for each of the five types of search systems are highlighted in Fig 2b. The subjective ratings given for the visual systems were consistently higher than those for Linear Search and Traditional Index for every characteristic. Furthermore, 93% of the participants concurred that the visual systems are either a major (12 out of 14) or minor (1 out of 14) improvement over traditional indexing. Comments such as "The relevance of pages can be shown easily in a graphical manner" and "These (The Visual Systems) are much faster and more dynamic – I don't have to plough through pages of indexes to find the one I want" further substantiate this improvement.

This data confirms our underlying hypothesis; namely, that colour and size both play a productive roll in the effective visualisation of keyword occurrences. Using a 7 point Likert scale, we asked the participants to rate the temperature colour system and the different sizes (tag clouds and graphs) for how well they illustrate the number of occurrences of a keyword on a particular page. The results from this yielded average results of 5 out of 7 and 6 out of 7 for colour and size respectively. The popularity of the visual systems are also backed up by the positive comments made by some of the participants. One for example, describes the graph system as "Very clear, big, bold and picture perfect" whereas another said "It's obvious which choices are the best and you can see them very quickly" with regards to the Tag Cloud. One participant even went as far as saying that the colour system was "common sense" and the size system "gives emphasis to the pages where there are more occurrences".

The analysis on the task timings confirmed that there is no statistical significance to suggest any differences in performance between the 3 visual systems. However, the post-study questionnaires have produced some interesting subjective results indicating a combination of both size and colour is the optimum method of visualisation. When asked which system they favour, all 14 of the participants selected either Tag Cloud (5/14) or Graph (9/14) and justified it with answers such as "It was easy to use and it appealed to me because I could visualise it" and "the colour and size of the page numbers make it easy to see".

## 5 Conclusions

This paper has explored the concept of custom index builders for use in digital document readers. It has described three unique visualisations designed to aid users in locating the most relevant sections in a document. These three approaches have utilised colour and size in an attempt to tap into existing cognitive mappings and provide a visual overview of keyword occurrences in a document. A provisional study of these implemented solutions concluded both quantitatively and subjectively that custom indexing is a large improvement over traditional linear searching. In addition to this, it also confirms that colour and size play a constructive role in their visualisation by proving that users favour them over traditional index systems.

## Acknowledgements

Jennifer Pearson is sponsored by Microsoft Research and Harold Thimbleby is a Royal Society-Leverhulme Trust Senior Research Fellow. We gratefully acknowledge their support. This research is supported by EPSRC Grant GR/S84798.

## References

1. Byrd, D.: A scrollbar-based visualization for document navigation. In: DL 1999: Fourth ACM conference on Digital libraries, pp. 122–129. ACM Press, New York (1999)
2. Harper, D.J., Koychev, I., Sun, Y.: Query-based document skimming: A user-centred evaluation of relevance profiling. In: Sebastiani, F. (ed.) ECIR 2003. LNCS, vol. 2633, pp. 377–392. Springer, Heidelberg (2003)
3. Loizides, F., Buchanan, G.R.: The myth of find: user behaviour and attitudes towards the basic search feature. In: Proc. JCDL 2008, pp. 48–51. ACM Press, New York (2008)
4. Ware, C.: Information Visualization: Perception for Design. Morgan Kaufmann, San Francisco (2004)