

Indicators of Content : the role of word clouds in the creation of summaries

by

Andrea Ennis

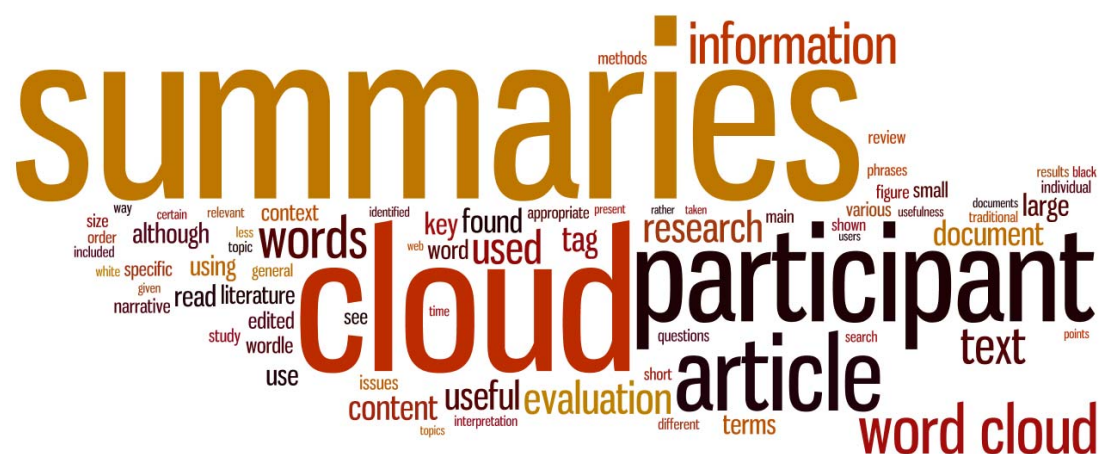
A Masters Dissertation, submitted in partial fulfilment of the requirements
for the award of Master of Science degree of Loughborough University.

September 2010

Supervisor: Dr Ann O'Brien
Department of Information Science

© A. Ennis, 2010

Abstract



Purpose

The purpose of the research was to explore the use of word clouds as indicators of document content, and to assess their use as summaries within an academic context.

Methodology

A literature review identified current uses for summaries, issues in their production and the use of Web 2.0 technologies as indicators of content. Research then took place with six participants who completed questionnaires to determine their initial impressions of summaries, before viewing various word cloud summaries and completing further questionnaires on the usefulness of these.

Findings

Word clouds were a useful indicator of document content, and helped participants to decide whether an article was relevant to their specific research topic. Large clouds were preferred by participants, to clouds with lower word count and edited content.

Impact

As clouds were found to be useful, and were quick to create while being relatively inexpensive, these could possibly be used as document summaries in place of, or supplementary to, traditional narrative summaries.

Research Implications

Future research could involve a larger number of participants, as this study used six volunteers, and word clouds as document summaries for people who speak English as a second language could also be considered.

Originality

Although the literature review revealed various uses for word clouds, clouds had not been studied as indicators of content for document selection.

Acknowledgements

Many thanks to my supervisor, Ann O'Brien for all her help and support

Contents

Abstract	ii
Acknowledgements	iv
List of figures	vi
List of tables	vii
Chapter 1: Introduction	1
1.1 Introduction	1
1.2 Aim	3
1.3 Objectives	3
1.4 Key Terms	3
Chapter 2: Literature Review	4
2.1 Traditional summaries	4
2.1.1 Automatic vs human summaries	4
2.1.2 Summary evaluation	5
2.2 Modern summaries	7
2.2.1 Recent developments in summary production	7
2.2.2 Web 2.0	9
2.3 Word clouds as summaries	10
2.3.1 Types of cloud	10
2.3.2 Cloud applications	11
2.3.3 Cloud evaluation	13
2.4 Conclusion	14
Chapter 3: Methods	15
3.1 Research methods general	15
3.2 Research design	16
3.3 Word cloud construction	21
3.3.1 Unedited	21
3.3.2 Edited	21
3.3.3 All clouds	22
3.4 Questionnaire planning	24
3.5 Pilot	26
3.6 Limitations of research	27
3.7 Ethics	29
Chapter 4: Findings and Discussion	30
4.1 Results	30
4.1.1 'Before' questionnaire	30

4.1.2 ‘During’ and ‘after’ questionnaires	32
4.1.3 General cloud comments	34
4.1.4 Specific cloud comments.....	34
4.1.5 Colour vs black & white	36
4.2 Discussion	37
Chapter 5: Conclusion	41
5.1 Range of current summaries.....	41
5.2 Current issues in the production of summaries.....	41
5.3 Use of Web 2.0 as indicator of content summaries	42
5.4 Appropriate situations	43
5.5 ‘Good’ summaries	43
5.6 Evaluate usefulness of summaries	43
5.7 Limitations.....	45
5.8 Recommendations for further research	45
5.9 Concluding remarks	46
Bibliography.....	47
Other Sources Consulted	51
Appendices.....	54
Appendix A - Participant recruitment	54
Appendix B – Participant session	55
Appendix C – Questionnaires	58
“Before” questionnaire.....	58
“During” questionnaire.....	60
“After” questionnaire – accompanying word cloud sheet	61
“After” questionnaire.....	62
Colour vs. black and white	65

List of figures

Figure 1: Flow chart to show research design.....	18
Figure 2: Screenshot of the first website page participants saw, with “search results”	20
Figure 3: Flow chart to show participant task flow.....	23
Figure 4: Number of responses for clouds most likely to encourage participants to read the corresponding article, and which cloud type was liked best.....	32
Figure 5: Percentage of each cloud participants said they had read.....	33
Figure 6: Percentage assessment of general impression of article content.....	33

Figure 7: Percentage assessment of understanding of position and argument of the article.....	33
Figure 8: Word cloud in which participant noted cited authors.....	35
Figure 9: Edited word and tag cloud including the terms 'pre-prints' and 'post-prints'	36
Figure 10: Word cloud showing phrase 'Myers Briggs Type Indicator'.....	38
Figure 11: Word clouds where the words 'guilt' and 'bond' were noted, suggesting desirable arguments.....	40

List of tables

Table 1: Cloud types and properties.....	25
Table 2: Scores for importance of specific summary components.....	31

Chapter 1: Introduction

1.1 Introduction

Summaries are used in a variety of situations, often for document retrieval, browsing and to ascertain the gist of a document. They are currently produced either by a human abstractor, or through utilisation of automatic summarisation programmes which statistically analyse text before presenting extracts as a summary (Spärck Jones 2007). Both of these methods tend to produce concise narratives that attempt to preserve the integrity of the original document. Their length ensures that they are a convenient and quick way to interpret and access information.

Various research programmes and projects exist to study document summarisation, emphasising their importance and relevance in an information rich environment. These projects focus on the different methods involved in the production of summaries, and on evaluation.

Academic papers make heavy use of summaries, it would be unusual to find an article without an accompanying summary or abstract to give the reader an outline of key topics and arguments. These are often produced by the authors themselves, or author summaries are further edited by professional abstractors for use in journal databases. Although traditionally, articles are represented by a paragraph of text giving an overview of the main arguments of the paper, some databases divide and present the summary under key headings, such as methodology and originality (as found on Emerald Insight¹), which allow details of specific areas of interest to be represented. Often, references or the headings included within the corresponding article will be made available to supplement the information included in the narrative summary (as found on ACM Portal² and Science Direct³).

Although widely used, such summaries have come under criticism for several reasons. Human summary production is expensive and takes time, and development of accurate automatic summary generators is a complex and again, time-consuming process. Issues with evaluation of narrative summaries, such as their subjective representation of an original text (Hirst 2007; Johnson 1995; Morris 2010; Zhan, et al 2009), and their lack of adaptation to the demands of small screen devices (Sweeney, et al 2008) are as yet to be resolved.

¹ Emerald Insight <<http://www.emeraldinsight.com/>>

² ACM Portal <<http://portal.acm.org/portal.cfm>>

³ Science Direct <<http://www.sciencedirect.com/>>

In response to these issues and by taking into consideration the usefulness of certain non-narrative summaries, such as those produced by search engines to display results, it is proposed that narratives may not always be necessary, and that word clouds, as an indicator of document content could be more appropriate in certain situations.

Word clouds are already being utilised as indicator of content summaries, most visibly by the media. A feature of the Obama/McCain US 2008 election was analysis of the blogs of both parties and of their speeches using word cloud visualisations (Schwenkler 2008). More recently, The Guardian newspaper made a cloud of a statement from Tiger Woods (Gibson 2010). Word clouds serve a purpose in media for being eye-catching, novel and for presenting key points in a quick, effective way. It is proposed that these representations could also be effective for presenting the content of academic articles.

To further assess use of word clouds and in order to explore the proposal that clouds may sometimes be more appropriate as article summaries, several research questions have been explored. What current range of summaries exist and what these are used for was established, and from this details of issues in production and use of summaries have been explored. Various Web 2.0 technologies were assessed, focusing especially on specific tools that could create word clouds, and situations where word clouds have been found to be appropriate were identified.

The majority of these questions were considered through the literature review, and the conclusions taken into consideration for the methodology. Further research questions explored here included establishing what users expected of a “good” summary, and assessment of how useful they found word cloud summaries in a set situation, in order to determine whether cloud indicators of content compared with narrative summaries.

From these research questions, the aim and objectives of the research were identified:

1.2 Aim

To explore the use of word clouds as indicators of document content, and assess whether these are useful as summaries in an academic context.

1.3 Objectives

- to identify the range of current summaries by means of a literature review
- to analyse current issues in the production of summaries
- to assess the use of some Web 2.0 technologies as indicator of content summaries
- to investigate situations where indicators of content are appropriate
- to determine user perceptions of “good” summaries
- to evaluate the usefulness of indicator of content summaries

1.4 Key Terms

Before continuing, it is important to define some of the key terms that will be used throughout proposal:

Gold standard The term used for, ‘human reference, model’ summaries (Spärck Jones 2007, p. 1454)

Indicators of content Summaries that provide only an indicator, rather than a detailed narrative explanation of a documents content

Web 2.0 Certain web based technologies that, ‘encourage participation, where all users of the web can create and add to the content available’ (Cosh et al 2008, p. 722)

Chapter 2: Literature Review

2.1 Traditional summaries

2.1.1 Automatic vs human summaries

Document summarisation, where the content of a source is reduced to a generalisation of the full text, has been studied for years. Summaries are used for document retrieval, skimming overviews and general browsing, and use determines the likely type and length of a summary. Professional human abstractors will take a document and summarise using their manual expertise, while automated methods have developed since the late 1950s.

Early automated attempts were dependent on statistical methods. Luhn (1958) studied themes such as term frequency and distribution. The increased interest in the field during and since the mid-1990s led to a huge number of papers and approaches being used, where currently, due to developments in semantic technologies, automatic summaries tend to be extractive, using statistical methods which rank sentences before extracting those it has statistically determined reflect the themes and main topics of the document. The developments to this point are brought together by Spärck Jones (2007), who identifies key themes in the literature up to the present day.

Various research programmes and projects, such as the DUC (Document Understanding Conference), SUMMAC (Summarisation Evaluation Conference) and NTCIR (National Institute for Informatics Test Collection for Information Retrieval), have studied document summarisation and evaluation. Their developments are summarised by Spärck Jones (2007) who outlines the various roadmaps for evaluation used by the DUC, which cover different types of summaries, evaluation methods, and single and multi-document summaries. It is worth noting that several programmes exist outside the programmes mentioned, including summary R&D (Research and Development) for Scandinavian languages, SweSum and ScandSum (Dalianis, et al 2003).

Spärck Jones' (2007) paper outlines two categories of summary; extractive, where general content is taken directly from the source document to create a shorter version, and non-extractive, where new sentences are created to be representative of the actual content (p. 1453).

Both extractive and non-extractive summaries can then be classed as either indicative or informative. Indicative summaries are used for reference and alerting,

giving a general overview of a text, where informative summaries can stand in place of a text (Mani, et al 2002).

2.1.2 Summary evaluation

Classifying summaries as extractive, non-extractive, indicative or informative leads to a wide variety of summaries being produced. Evaluation methods differ accordingly, dependant on the individual summary.

Summaries can be evaluated intrinsically and extrinsically. Intrinsic evaluation relates to the internal construction of the summary itself, and is mainly dependent on the summary being linguistically well-formed, with concepts genuinely reflected, and should compare to a “gold standard” (human summary). Extrinsic evaluation relates to external use, and, ‘tests the summarisation dependent on how it affects the completion of some other task’ (Mani et al, 2002, p. 44). Benefits of extrinsic evaluation are that it tests summaries in real world situations, and as such can detect differences between summary systems (Over, et al 2007, p. 1508).

Evaluation and research was initially focused on single document summaries, but has evolved to include the summarisation of multiple documents, often known as multi-document summarisation (Wei, et al 2009). Newsblaster⁴ used multi-document summarisation to aggregate news sources to produce a single summary of one news story. The same complex evaluation criteria apply, often leading to the production of relatively lengthy narratives in order to be accepted as a decent summary.

The evaluation criteria and standards enforced on automatic summaries make evaluation problematic. Interpretation of what is meant by “proper” discourse, and comparison to a gold standard are not objective criteria in themselves. As part of addressing evaluation issues, most DUC summaries are done using news articles, as the way these are written make finding a more appropriate target challenging (Over, et al 2007, p. 1513).

Where automatic summaries strive to provide an ideal summary, a tautology is created, as there is no ideal situation in which they can be applied (Johnson 1995). A gold standard summary will still be a human abstractors interpretation of a document with a certain audience and purpose in mind. As such, to generate an ideal automatic abstract, there would need to be an understanding of what particular

⁴ Newsblaster <<http://newsblaster.cs.columbia.edu/>>

parts of abstracts were useful or not in various situations, information that is unlikely to be known (Johnson 1995, p. 34).

Other studies identify the problem of evaluation as having arisen from the assumption that a document or block of text can be statistically evaluated. Statistical extraction of text assumes that the meaning of a document objectively “lives” in the text, and that any reduction of this to a summary can include the meaning. It denies any writer/reader interpretation, which was the previously popular assumption within computational linguistics (Hirst 2007).

One such interpretation is of ‘lexical cohesion’ (Morris 2010, p. 141). Individuals interpret lexical cohesion in different ways, and as such various interpretations and individual understanding can apply to one text (which could explain why gold standard human summaries differ).

Where Johnson (1995) was focused on the situation the summary was being used in, and how automatic summaries could never determine these, Hirst (2007) and Morris (2010) similarly bring attention to the denial of the individual. The specificity of a person in a certain situation is important to understanding, and statistical methods ignore this specificity, reflecting only one version. This is true, however, for human summaries as well, as only the interpretation of the person summarising is presented.

After discovering bias in human summaries, Zhan, et al (2009) studied human multi-document summaries and length, and distinguished between macro (factual, “bones” of the document) and micro (supporting detail) information that appeared in the final summaries.

They found that in very short human summaries (<50 words), the ‘macrostructure’ (p. 58) of the documents were not iterated, but when authors were given a larger word count, although the summaries still contained different words, they started to adopt a similar structure. Micro level information was included in addition to this basic structure depending subjectively on the authors’ experience.

This could have implications for both Hirst (2007) and Morris (2010), as it seems to suggest an objective meaning and interpretation can be taken from the text, however it is interesting that this is dependent on length. As mentioned before, to satisfy traditional evaluation criteria summaries tend to be relatively lengthy, but Zhan et al (2009) determine that between 50 and 200 words is where the

'macrostructure' of a summary becomes clear, with micro information creeping in at 200 words.

The DUC has studied short summaries, of ≤ 75 bytes, as this was seen as the typical size for electronic data (Over, et al 2007, p. 1512). Shorter summaries were given different evaluation criteria to their longer counterparts, and could include non-standard grammar, while still being assessed as decent summaries. Linguistic cohesion is demanded of longer summaries to prevent misleading or nonsensical lines of text (Over, et al 2007, p. 1512).

Short summaries evaluated by the DUC were evaluated intrinsically, as are most DUC evaluations (Over, et al 2007, p. 1514). Johnson (1995) suggests that to overcome the tautology of "ideal" summary and situation, evaluation methods need to be extrinsic. Instead of using gold standards, success should be measured by how useful a summary was in relation to users needs (Johnson 1995, p. 35).

2.2 Modern summaries

2.2.1 Recent developments in summary production

Evaluation issues aside, automatic summaries continue to develop, and have many web-based applications. Arguably the least comprehensive short summaries are found for internet search results, where an extractive, indicative summary of information that you would find on the page is presented. These are categorised by Li and Chen (2010) as 'snippet extraction', where, 'one or multiple parts of a document that contain as much of a query as possible' (p. 378), are displayed.

These basic summaries are used by various search engines; Google⁵ returns snippets, and search engines designed for content management systems such as Microsoft Sharepoint's Fast Search Server⁶ identify fragments containing key words. A step ahead of key word selection, Coveo⁷ returns the most significant sentence from the document being searched. Similarly, a MEDLINE trial focused on short, one-sentence summaries, taking a key sentence from a MEDLINE abstract to be indicative of an articles content (Ruch, et al 2007). Keywords are largely used within MEDLINE to get a "gist" of an article, and it was assumed that there would be a similar use and application for sentences. More so than keywords, Ruch found that defining key sentences was more complicated than key words as key

⁵ Google <<http://www.google.com/>>

⁶ Sharepoint Fast Search Server <<http://technet.microsoft.com/en-us/evalcenter/ee424282.aspx>>

⁷ Coveo <<http://www.coveo.com/en/products/platform/technical-overview/advanced-technology>>

sentences are dependent on reader interpretation. This is not the only reason simple extractive summaries are not always appropriate, as although extracted from a full text document or website, the summary does not necessarily reflect the entire content of the page, it only documents exactly what matches a search.

Sensebot⁸ uses text-mining and multi-document summarisation to give a short automatic summary of site content, which provides much more of an overview than basic snippet or sentence extraction, by including a general overview of whole site information.

As short summaries, these web based applications minimise subjective information and interpretation as previously with Zhan, et al (2009) and macro and micro information. Where these have been shown as useful for searching, by presenting succinct information these can also be utilised in response to display restrictions on small devices. With an increase in technologies such as smart phones and tablets, demand for condensed summaries have increased, although there was previously demand for this from visually impaired communities (Harper and Patel 2005).

In addition to previously mentioned display restrictions, mobile access differs as users demand and use information in different situations to that of traditional desk/computer set-ups and are possibly paying data charges for everything downloaded. For some, automatic summaries are an ideal solution to many mobile data issues. Several diverse projects have recently been developed utilising shorter summaries.

Document cards to hold short summaries with key text and figures were developed for display on small devices (Strobelt, et al 2009). Summarisation, rather than being held as an accurate representation of the full documents content (as statistical extraction was previously discussed as being assumed to be, where the full meaning of the document can be assumed from the document alone without writer/reader interpretation), is a, 'lossy compression and requires a decision of what can be preserved and what has to be excluded' (Strobelt, et al 2009, p. 1146).

Hierarchical text summarisation has been used to create short summaries giving only novel information (Sweeney, et al 2008). A, 'top level summary' (p. 668), gives basic, minimum information about the document, then each hierarchy adds more detail until finally the full text is revealed. This direct, systematic approach is well-suited to smaller screen size restrictions.

⁸ Sensebot <<http://www.sensebot.net/>>

The hierarchical text summarisation used by Sweeney et al (2008) is also presented as an alternative to linear summaries (Yang and Wang 2008). Where the majority, if not all automatic summaries present sentences in their original order, they present these as a sequence, while ignoring any hierarchical structure (p. 897). In presenting hierarchical, 'fractals' (p. 887), divided up using the original headings, the summaries cover a wider range of information by representing each section of a document (p. 901).

These projects use a variety of human and automated approaches to summarisation. All of them agree that some parts of a text are more "important" than others; that information will be lost through compression (Strobelt, et al 2009), "facts" are more weighted than novel information (Sweeney, et al 2008), and headings provide a set framework for understanding (Yang and Wang 2008). No matter how 'lossy' (Strobelt, et al 2008, p. 1146), the summarisation has been, all the studies were proved useful by user studies.

Innovation to meet demands of device screen size and information retrieval has led to the development of shorter and non-linear summaries. Unlike traditional narrative summaries, short summaries do not necessarily need to meet standards of linguistic cohesion (Over, et al 2007, p.1512), however, narrative can give context to a non-specialist user (Veel 2009, p. 92). Fuzzy concepts and metaphors do not necessarily make complete sense when taken out of a narrative context, as when MEDLINE sentences were extracted (Ruch, et al 2007), sometimes random sentences can make less sense. Short and non-linear summaries will not necessarily give the complete context, however a summary given at the start of a piece of text has been shown to increase the speed of reading of the text. When the summary was given at the end, although questions on understanding were answered with approximately the same accuracy, users were slower to read the text (Giora 1985, pp. 128-131). Short summaries can still be useful outlining main linguistic features of a text, even if the subtleties may be lacking.

2.2.2 Web 2.0

Web 2.0 technologies offer many semantic tools that support the idea of the occasional redundancy of lengthy, detailed summaries, instead offering basic topical information that still informs. On social networking sites 'microblogging' allows short summaries rather than an entire blog entry to update on user interests.

Twitter⁹ and Tumblr¹⁰ are two microblogging sites, where a user enters a short summary (Twitter has a 140 character limit) of their thoughts, rather than writing a full blog entry. Mainly a tool for social networking, this could have possible summary implications, as often on table of content alerts only the title of an article is given and these are not always indicative of content. Very short summaries were studied by the DUC in 2003. Some of the summaries were around 10 words, and these were still assessed positively (Alfonseca and Rodriguez 2004). A short Twitter-style alert may be useful in certain situations.

Not all tools are entirely text based. Mycrocosm¹¹, although updated by text, creates a graph of the users' topic choice. Part way between text and graphical display are word clouds, which although display text, only as disassociated words and terms, and in a random graphical display.

2.3 Word clouds as summaries

2.3.1 Types of cloud

There are several variations of word clouds although generally, the term “word cloud” relates to a cloud created from statistical analysis of word frequency, variations include using either one or two term frequencies (ManyEyes)¹², the shape of the cloud (Tagxedo¹³), and where some are basic visualisations, others provide additional information such as pop-up details of frequency and where in the text the word was found (ManyEyes). Various terminology and names given to different clouds also relate to different aspects being prioritised.

Tag clouds, as one variation, usually represent tags that have been added to a document, instead of a statistical analysis of the text. Size of tag represents how many times a document has been tagged with a particular term. Tag clouds often have the added functionality of being interactive, if you click on a specific term you can be taken to a list of items tagged with that particular term.

Tree clouds (Gambette and Véronis 2009) have been used to detail a term and the terms that are then directly related to them (Freedman 2010). The “branches” of the tree reflect both the distance and the path of how terms are related. Clouds have also been utilised as internet category maps, which have been designed to give an overview of the entire content of a site (Yang, et al 2003). Other interpretations use

⁹ Twitter <<http://twitter.com/>>

¹⁰ Tumblr <<http://www.tumblr.com/>>

¹¹ Mycrocosm <<http://mycro.media.mit.edu/>>

¹² ManyEyes <<http://manyeyes.alphaworks.ibm.com/manyeyes/>>

¹³ Tagxedo <<http://www.tagxedo.com/>>

traditional word clouds, but then map these onto separate layouts to add context. Clouds have been mapped onto layouts reflecting date and time (Cui, et al 2010), to provide further context to the words in the cloud, and also onto GIS (Geographical Information Systems) to add qualitative information to quantitative data, in an attempt to broaden understanding (Cidell 2010). Although now incredibly differentiated, a detailed history of clouds is given by Viégas and Wattenberg (2008), who trace use of various word clouds back nearly 100 years.

This visualisation approach of word clouds, representing by word size the frequency of a specific term, has mainly been utilised to present summaries and overviews of documents and topics. One study however, has been identified that utilises the visualisation approach within the context of the document. Rather than displaying the resized words as a cloud, the document narrative was preserved and resized words left in situ, the text size changing dependant on frequency of the word (Gottron 2009). This is an unusual use of the visualisation approach, however did help users identify the topic of the document more quickly than with a document with a continuous text size.

2.3.2 Cloud applications

As a Web 2.0 tool, clouds are mainly used in social networking. It is possible to generate clouds of status updates from various social networking sites (Twitter, Facebook¹⁴, Tumblr), and tag clouds are widely used by these sites as a way to access information related to a specific tag. Possibly partly due to this social use and interpretation, there is very little literature that addresses word cloud use as a possible summary for a document.

In the face of detailed, lengthy and regular DDBJ (DNA Database of Japan) releases, word clouds were used to provide a quick overview of the main themes of each document (Kaminuma, et al 2009 p. 4), saving time and effort reading complicated releases.

Similarly, clouds are used in academic papers to provide broad topic overviews. Rather than an extensive literature review, which may not always be appropriate, large or multiple academic documents were summarised in a cloud (Cooper, et al 2009; Lester and Robinson 2009).

These general overviews give an at a glance visualisation of otherwise time consuming reading, and by their compact nature, clouds have also been used for

¹⁴ Facebook <<http://www.facebook.com>>

comparisons. The GMC (General Medical Council) releases a yearly document entitled “Good Medical Practice”, and clouds were created from the text of these documents (Gill and Griffin 2010) to compare the changing, ‘meanings and leanings’ (p. 32) of the GMC. Again, as the documents were large, word clouds were an appropriate medium as they saved time and more easily distinguished implicit ideas from the text.

The media utilises word clouds, often as representations of speeches that would be either too long or irrelevant to print in full. The Guardian used a word cloud of a speech by Tiger Woods (Gibson 2010) as the accompanying picture to a story, and a large number of word clouds were generated during the US 2008 election, comparing speeches and the blogs of each presidential candidate (Schwenkler 2008). One of the main stories to come out of this analysis, was that the name ‘Obama’ was prominent in both parties’ blogs, however ‘McCain’, only in his own. In addition to providing an overview, the clouds produced were carefully scrutinised and compared to reach this insight. Word clouds can have a useful application that would otherwise be difficult to gauge from the full text.

A word cloud generator and system to upload, create and compare clouds called ‘Scriptcloud’ (McKie 2007) was set up for screenplay comparisons. Professionals are hired at great cost to analyse scripts and advise on how to improve plots and themes. Creation of the website allowed large screenplay script themes to be analysed quickly by amateur script readers.

Other literature on word clouds shows uses information professionals have found for them. Consideration has been made as to cloud use in indexing (Burmeister 2010), both as a more aesthetically pleasing list of indexing terms than a list, and as a representation of indexing terms for a collection of documents. Clouds here would be used as content indicators, rather than their more widely known use as access points to blog topics.

Information professionals have mainly used word clouds alongside searching and browsing. Some of the benefits discovered through utilising word clouds in this way, and seeing how users react to them could also be relevant to summarisation.

Many studies have found that word clouds used alongside traditional catalogue searching encouraged navigation (Mayfield, et al 2008; Olson 2007; Seifert, et al 2008). Word clouds expose controlled vocabularies that are not necessarily evident to the user (Olson 2007) and similarly, searches using word clouds reveal a, ‘link to

relevant subsets' (Seifert, et al 2008, p. 17; Yang, et al 2003) that may have previously been unrealised.

As a practical demonstration of how people use clouds (specifically tag clouds), Kuo, et al (2007) created a tag cloud (PubCloud), using words taken from the abstracts of articles on PubMed. Users were given questions and had to give an answer after searching either PubCloud or PubMed. They found descriptive questions favoured PubCloud, but relational questions took nearly twice as long than with the traditional search engine. Where descriptive information is key to document understanding, word clouds give a clear picture of the main descriptors, however may be less suitable for documents with abstract relational concepts. This has also been emphasised elsewhere (Burmeister 2010), as word clouds were suggested to present the topics inherent in a document very well, with arguments and direction more elusive.

2.3.3 Cloud evaluation

Outside of catalogue and information searching applications, evaluation of word clouds has mainly taken place through interviews (Hearst and Rosner 2008; Viégas, et al 2009). Extrinsic evaluation focused on general impressions formed by users when shown clouds (gisting) (Rivandeniera, et al 2007) and intrinsic evaluation on what word clouds added to, 'exploratory data analysis' (Cidell 2010, p. 5). Several general issues have been raised through these studies, generally addressing user issues, information distribution and scrutiny and comparison.

Heavy users of Wordle¹⁵, a word cloud generator, have varying levels of comprehension of what different visualisations (text size, colour) actually represent (Viégas, et al 2009). Even when users comprehend these visualisations, the distribution of words can be misleading; too many words can be less intuitive (Yang, et al 2003), and Western reading patterns can place greater importance on the top left corner of the cloud (Rivandeniera, et al 2007; Hearst and Rosner 2008), and similar or related words can end up far apart (Hearst and Rosner 2008). For scrutiny and comparison, as word clouds are independently sized and relational (Cidell 2010), making visual comparisons can be misleading (Cidell 2010; Hearst and Rosner 2008).

The majority of these criticisms can be addressed either through use of a variation of a word cloud; distance between related words can be visualised in a word tree

¹⁵ Wordle <<http://www.wordle.net/>>

(Freedman 2010; Gambette and Véronis 2009), and tag clouds can offer a degree of abstraction for complex concepts (Cosh et al 2008). Distance between terms has also been put forward as a positive reason for using word clouds, a randomised pattern of words, even if unrelated can lead to discovery (Hearst and Rosner 2008), and although clouds are relational to the size of the source document, they were specifically chosen and useful for comparison of GMC documents (Gill and Griffin 2010).

The major issue for word clouds that has come out of extrinsic research is that when comparing word clouds to a list of words in frequency order, the lists are shown to provide a better 'gist' of the text (Rivandeira, et al 2007, p. 998) when accuracy was studied. Users were shown a cloud for 30 seconds and then had to identify categories contained in them, receiving a point for each one correctly identified. Word size is identified in each test as contributing to user recall and identification, and here it is implied that word size highlights main categories, however distracts from smaller categories.

2.4 Conclusion

Summaries have been used for searching, cataloguing and overviews for a long time, and more recently the internet and the demands of information retrieval have shown that traditional summaries, where a linguistically well-formed narrative is demanded, may be redundant in certain situations. Evaluation of traditionally formed summaries has been criticised as user interpretation is not included, and narrative structures are arguably subjective.

As screen size for mobile devices decreases, and Web 2.0 technologies become more prominent, it seems traditional summaries are not always necessary, and in information intensive environments, that indicators of content such as word clouds could be more appropriate. Due to the non-narrative structure of such summaries, traditional evaluation criteria will not be relevant, and other evaluation methods will need to be found.

Chapter 3: Methods

The literature review showed that evaluation of summaries has been problematic, with various methods used. This chapter outlines a method, in the context of the aim of this research that explores the use of word clouds as non-narrative indicators of document content, and assesses whether these are useful as summaries when used for article selection. Traditional evaluation issues are challenged by placing evaluation on the participant and their personal opinion of how useful a summary was to their information seeking work.

3.1 Research methods general

Summary evaluation has mainly taken place using quantitative research methods. Summaries have been evaluated intrinsically and extrinsically, and intrinsic evaluation and methods such as statistical evaluation have been favoured by the DUC. The subjective evaluation criteria this often includes however, such as opinions on what constitutes lexical cohesion, have been criticised, as outlined in the literature review (Johnson 1995, Hirst 2007, Morris 2010).

Extrinsic evaluation has been utilised to evaluate shorter summaries, with studies measuring the performance of a summary against selected tasks such as identification of relevant documents and categorisation (Mani, et al 2002), and metrics such as time, accuracy (Mani, et al 2002) and precision (Yang and Wang, 2008). Such metrics measure quantitatively, calculating how “correct” a user is at identifying documents from summaries, and how quickly they can do it, measured against set metrics.

Word cloud evaluation has mainly been qualitative, taking place through interviews and assessment of clouds by individuals (Hearst and Rosner 2008, Viégas, et al 2009). The quantitative evaluations that have taken place have generally focused on the task of categorisation, with evaluation taking place against set “correct” responses (Rivandeneira, et al 2007). This shift to qualitative methods is possibly due to the fact that word clouds do not emulate the conditions demanded by quantitative methods. The metrics used measure by criteria such as the number of sentences accepted by a user (Yang and Wang 2008), and such metrics are only valid for summaries that present a narrative of document content.

These quantitative methods have also attracted criticism as to their usefulness in assessment. By defining set parameters such as categories (Rivandeneira, et al 2007), relevance (Sweeney, et al 2008) and precision (Yang and Wang 2008), an “ideal” summary is created and assumed to exist within these boundaries.

Summaries can only be categorised and identified in one “correct” way, no other opinion is valid. This creation of a perfect scenario, where each summary can be assessed in the same way in any given context, has been said not to exist (Johnson 1995).

As such, although setting parameters is one way to assess usefulness, as part of a quantitative evaluation method it does not take into account the personal opinion of the user, and how effective they have found the summary. User opinion of effectiveness is judged from the results of their answers to pre-determined, researcher-set purpose questions.

Due to the issues with quantitative methods in assessing summaries, qualitative methods have primarily been used in this research. It was proposed that a heuristic evaluation of word clouds would satisfy some of the criticisms levelled at quantitative methods. Here a participant offered his/her opinion on how effective the summary was found in the frame of his/her own research and information needs, rather than in circumstances that were dictated. Their assessment was of the “gist” that they gathered from the summary and whether it was enough to go ahead and read the full text was evaluated.

Task selection and metrics for evaluation are problematic in quantitative models. Using mainly qualitative methods ensures task selection is relevant and specific to individuals as they are the ones most aware of their own information needs. Heuristic evaluation selected tasks specifically in this way in order to reflect realistic scenarios and metrics were then participant-defined and evaluated. Success was measured in reference to how useful the summary was to a participant.

3.2 Research design

Participants were chosen randomly from Information Science postgraduate students at Loughborough University. This group was chosen as access was feasible, they use summaries frequently and are familiar with traditional summary formats, and they were also known to have produced research proposals that were being developed into dissertations. A pilot was run with one participant to test the format and feasibility of the research design (see Figure 1 for research design flow chart).

Participants were approached by email (see Appendix A), which explained that research was being carried out into summaries. Word clouds were not specifically mentioned in an attempt to reduce bias and pre-judgement of outcomes. Each of the six participants who agreed to take part was asked to explain their current

dissertation topic and to provide the bibliography from their original research proposal. These were used to source articles relevant to their dissertation topic, and for word clouds to be created from these. Explanation of the topics took place either as an informal interview or via email exchange, in order to understand the key terms and concepts of each participant's research.

All participants were asked to indicate the stage they felt they had reached with writing their dissertation, information which was noted but kept anonymous. It was anticipated that the participants who had progressed further with their research were likely to be more aware of their topic and have more precise information needs. There was also the likely scenario that they would have developed their literature review and identified a larger number of relevant articles than were listed in their original research proposal bibliography. As such they may already have found and been familiar with the articles represented in the study. Knowledge of the full text may influence whether the article was accepted or rejected. Efforts were made both to avoid obvious articles and to target articles that although relevant, were slightly more unusual in topic.

In choosing articles an attempt was made to reflect both descriptive and more abstract, analytical articles, so summaries covered a range of styles, as the literature indicated that word clouds may more clearly represent descriptive articles than those with complex concepts (Kuo, et al 2007). Presenting a selection ensured that even participants with abstract topics saw descriptive articles, and those with descriptive topics had a fair representation of abstract articles.

The participant's explanation of topic was used alongside the research proposal bibliography to identify key search terms. Searches were undertaken to identify articles on each topic, using the bibliography both to create a word cloud using the title and abstract of articles to help identify further search terms and also to eliminate articles that had already been read. The documents were in HTML or some readable format other than Acrobat files, as these were processed by a word cloud generator, which often accepted only plain text.

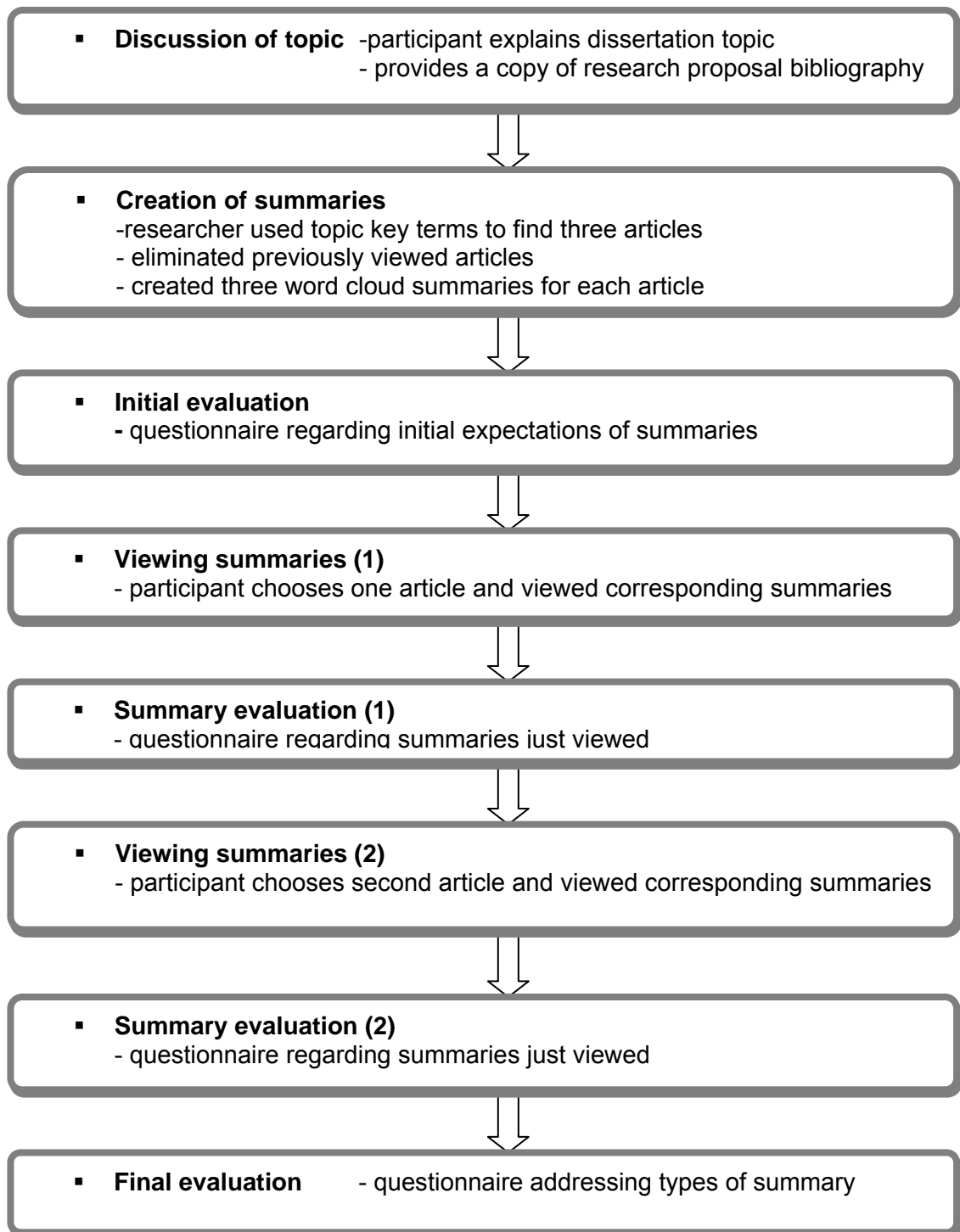


Figure 1: Flow chart to show research design

Three articles were identified for each participant's topic, culminating in 18 articles for 6 participants. Summaries of the articles were then produced. Each of these articles was provided with a corresponding narrative summary, however this was not shown to participants as they were already familiar with such summaries, and as such the usefulness of such summaries was assumed established.

Three word cloud summaries were produced for each article. Although several types of word cloud were identified in the literature review, due to copyright restrictions and access to cloud generators, only two generators were used, Wordle and TagCrowd. Wordle creates straightforward word clouds, representing word frequency by text size. It was used to generate the first and second summaries, the first from unedited text and the second from edited text. The third summary was generated by TagCrowd, which generates a tag cloud in alphabetical order, again from edited text. In the two instances where text was edited, this was done to remove assumed redundant information (see Table 1 for cloud outline).

A web site was designed to look as if articles had been retrieved from a search on the LISA database, as participants were all familiar with the format of search results using this database. This helped keep people focused on the task of choosing articles as if for their dissertation. Each participant had a separate site, with all three of their "results" showing title and author (see Figure 2).

Participants reviewed their own results individually, in one-to-one sessions. Before moving onto the summaries, each participant completed a questionnaire (see Appendix C), asking their opinion on what made summaries good or useful and what kind of features they looked for in a summary. What they would see on screen was explained; that they were to be presented with three articles that had come up in a search in relation to their dissertation topic and three summaries for each one. They were advised that they needed to select two of the three articles, and that that these summaries might be different to traditional summaries, and that they needed to consider whether they would or would not go on to read the corresponding full text, and what lead to this decision. Only two articles of the three were selected as participants then viewed each type of summary twice, without being overloaded with information.

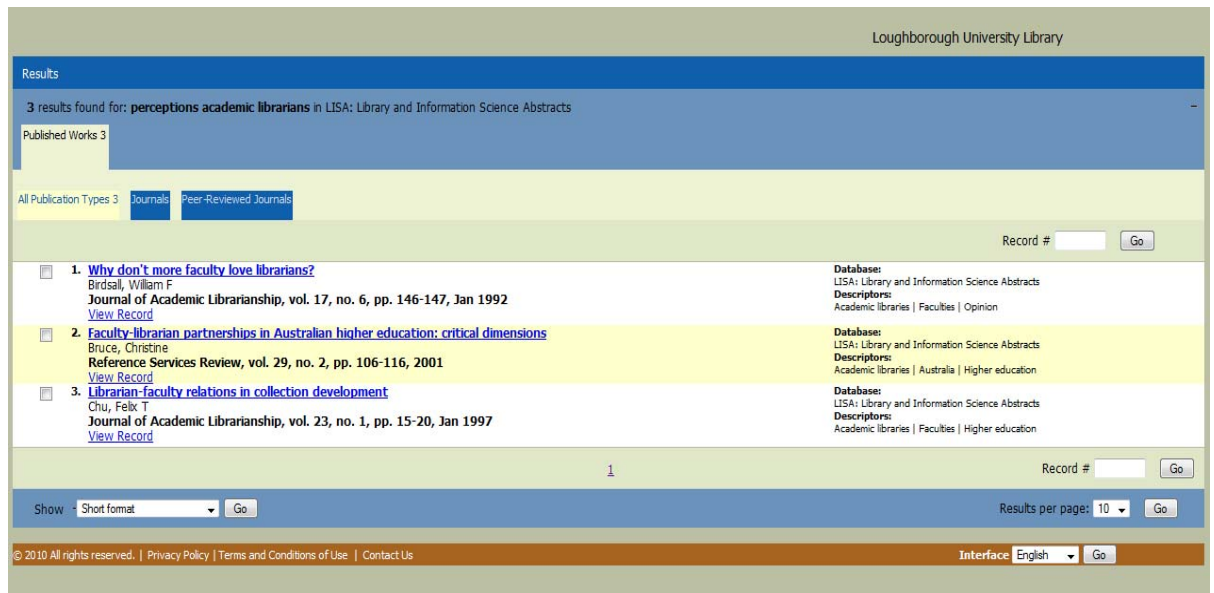


Figure 2: Screenshot of the first website page participants saw, with “search results”

Each participant viewed their results page, and chose their first title. Clicking on the title lead to the first summary, randomly selected from the three. When a decision was made whether to accept or reject, they clicked a Next button on the web page, which lead to a practically blank page with only text to click to the next summary. This summary, followed by blank page was repeated until all three summaries for one article had been seen, the site then returned to the original results page (see Appendix B for screen flow example).

After viewing a set of summaries for one article, the participant completed a short survey on whether they would read the article, and which summary would convince them to do this. This was repeated for both articles.

After viewing all summaries for both articles, the participant then answered a series of questions on the summaries seen; detailing their impressions of the different types of summary (see Figure 3 for outline of participant task flow).

After answering all questionnaires each participant was presented with a printed version of one of the clouds they had viewed (see Appendix C). This print-out showed the colour version of the cloud (already viewed), and the same cloud in black and white. Participants were then asked whether the colour affected their assessment of the summary at all.

Summaries were presented in a random order as common in the literature, in order to minimise bias (Yang and Wang 2008), and to control inference from previous summaries as much as possible. Whichever summary was shown first, the “gist” of

the article from that summary would have remained for the following summaries. Presenting the summaries in a random order and displaying a blank page between summaries prevented as much as possible preference for and inference from previous summaries.

Following this method provided data on what participants believed was useful in a summary and what they accepted/rejected and why. This contributed to being able to assess whether word clouds, as non-narrative structures were sometimes appropriate indicators of content in place of narrative summaries.

3.3 Word cloud construction

Participants were shown three word clouds. Two had been generated by Wordle, which displays words in a random order in a tag cloud format. Although tag clouds often have added functionality in that if you click a term, you can then be taken to the full text with the tag terms highlighted, this functionality was not available to be used. TagCrowd was chosen for its tag cloud alphabetical and linear layout.

Each cloud was created from either unedited or edited text (see Table 1 for cloud outline).

3.3.1 Unedited

Large clouds were generated using Wordle from unedited text, taking the entire article (not including bibliography) as the text to be entered into the cloud generator. These all displayed 100 words per cloud. The default display for Wordle is 150 words, however the pilot showed these clouds to be illegible.

3.3.2 Edited

Small clouds and tag clouds were generated using Wordle and Tagcrowd respectively, and used edited text, as some of the information was considered as possibly redundant and distracting. Text was edited to remove references where Harvard citation had been used within the text, and synonyms were amalgamated into one word, in order to display more accurately the prominence of a term. Some were considered best left in the cloud, as they were thought to display a useful breadth of terminology that may not have been previously recognised by the participant. Phrases were edited to appear together in the cloud (i.e. "information literacy") to give a clearer representation of present terms.

Removal of plurals is done automatically in TagCrowd, which displays the most used term to represent the word count of all the associated terms. This was done manually in Wordle. Removal of plurals is important in both, as it prevented

unnecessary repetition, which would be negative in any summary, and also allowed creation of a fair size representation of a term, counting singular and plural separately would have resulted in two smaller words rather than one large one.

Common English words are removed automatically in both Wordle and TagCrowd. TagCrowd allows use of a custom stop list of further words to be ignored and Wordle allows removal of individual terms. These capabilities were used in some articles as appropriate.

Small and tag clouds with edited text displayed any number of words from 30 to 70. This was dependent on the individual article, taking into consideration factors such as article length, and specific terms that were determined as relevant, that disappeared at certain word counts.

3.3.3 All clouds

Aesthetically, editing applied to all clouds, dependent on the generator. Large and small clouds generated by Wordle were all given a horizontal layout and a colour scheme was chosen to be standard across all clouds in colours that would still be viewable by colour-blind users. Including a black and white cloud was considered, as any print-out of a summary is likely to be done in black and white, however it was considered beyond the scope of this research to include a black and white cloud (see Limitations 3.6). This was introduced separately to the original summaries to gauge opinion.

Font is randomly assigned in Wordle and this was left as long as it was legible, if not this was changed to a basic font.

For small edited clouds, as Wordle distinguishes between capital and lower-case letters, these were all changed to lower case using the in-built functionality of the programme.

TagCrowd generated clouds were produced as HTML code and then edited using CSS to improve layout features such as font, padding and word spacing, as most of the literature discussing word clouds was focused on the presentation and usability of visualisations (Hearst and Rosner 2008; Cui, et al 2010). Other choices as described for Wordle were already managed by the programme itself.

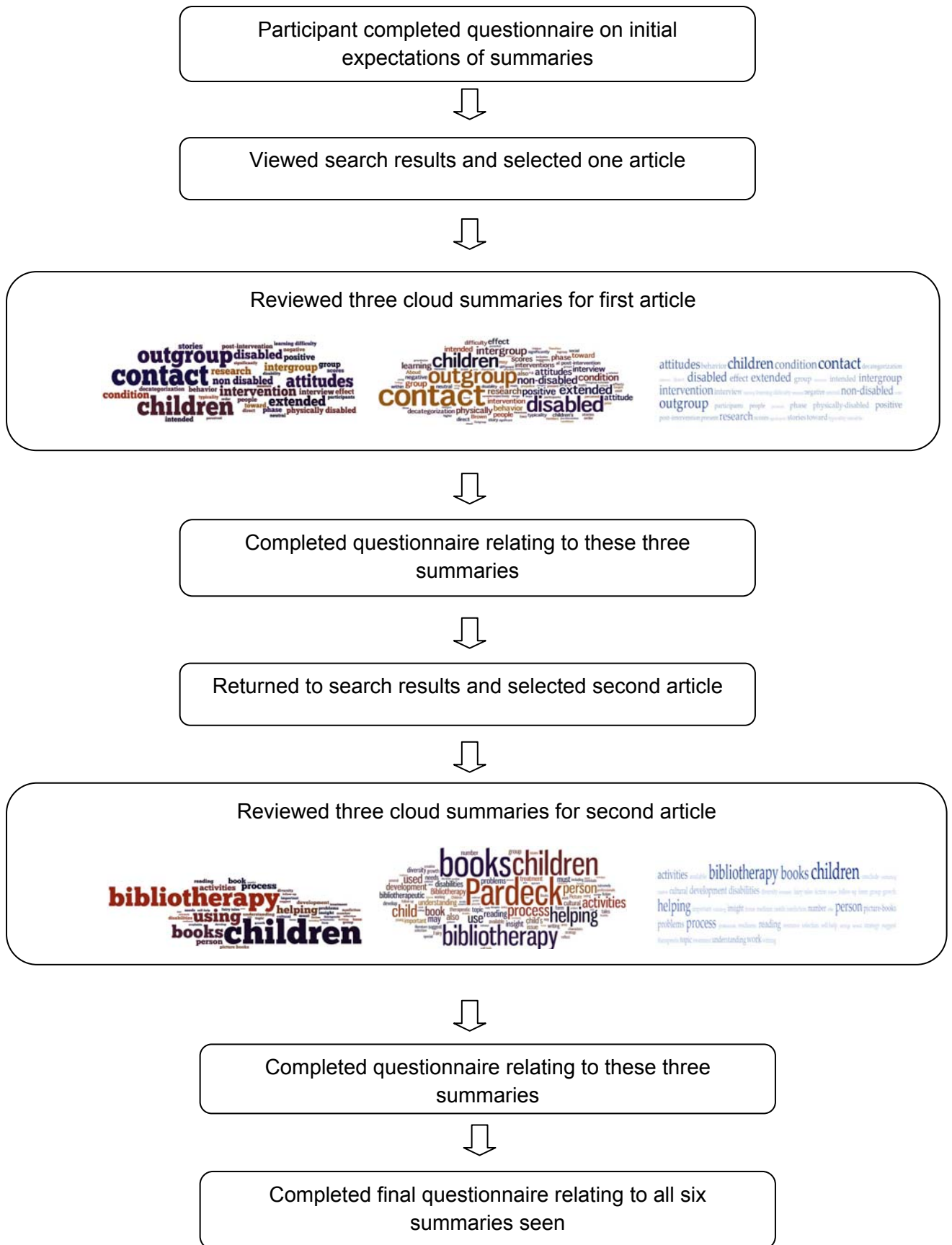


Figure 3: Flow chart to show participant task flow

Unedited clouds were very quick to generate, where edited clouds took approximately an hour per article.

3.4 Questionnaire planning

Participants completed four questionnaires (see Appendix C); one before seeing any summaries, one after each set of summaries for an article and one after reviewing all summaries. Questionnaires were chosen as an effective way of gathering opinion and experience. They were all self-administered, so the participant recorded their own responses.




Questionnaires with some open questions were chosen over methods such as an interview, as the aim of the research was to assess individual usefulness of the separate summaries. A standardised scale was needed to ensure general consistency in feedback, making questionnaires more feasible than interviews, as different vocabularies would have made analysis incredibly challenging. Interviews would have been more suitable for an in-depth description of how summaries were specifically useful for a certain purpose, where analysis of a general feeling of usefulness is more appropriate to summaries that are indicators of contents, where people would not necessarily be able to describe specific details of meeting purpose.

The first questionnaire was completed before participants had seen the summaries. They were briefed on the task they were about to complete; assessing articles for relevance using various summaries, and that the questionnaire was to gather their initial opinions on summaries. It aimed to discover what participants consider the components of a “good” summary, in the frame of researching their specific dissertation topic. Stages to answering this included exploring whether participants do find summaries useful and what criteria they were using to assess summaries.

These stages were measured with a five point Likert rating scale. Some questions included a scale from strongly disagree to strongly agree with various statements regarding the previous stages. Using a rating scale assured consistency in gauging which criteria participants were using to assess usefulness.

An open-ended question was included at the end of the questionnaire to establish whether participants took into account other criteria that had not been put forward.

Table 1: Cloud types and properties

Name of cloud	Created by	Text	Word count
<p>Small</p> 	Wordle	Edited	Determined by article, 30-70 words
<p>Large</p> 	Wordle	Un-edited	100
<p>Tag</p> 	TagCrowd	Edited	Determined by article, 30-70 words

The second and third questionnaires were answered after each set of summaries. Each article had three summaries and after the third had been viewed, this questionnaire was completed to assess whether the participant would go ahead to read the article, which summary convinced them to do this and which was their preferred summary.

The final questionnaire was completed after a participant had viewed both articles and the corresponding summaries. This assessed general impressions gathered from the summaries, what features made summaries useful, and what aspects contributed to this. Rating scales and open-ended questions were used to explore these stages.

A brief interview was conducted after participants viewed a black and white version of a cloud they had seen. This was seen as more appropriate than introducing a fifth questionnaire, that would possibly demand too much of a participant, and also because general, individual impressions were being gauged that were better addressed through an interview.

3.5 Pilot

A pilot was undertaken to test the research design. This involved one participant, who was not part of the participant group for the study. This pilot followed the research design (see Figure 1).

The pilot showed that the word clouds used on the test site were sometimes hard to read, with smaller words becoming illegible. The clouds had been generated and then saved as JPEG files, before being resized using HTML.

To make the clouds more legible, three steps were taken. The first was to change the word count of the unedited word cloud. This was shown by the pilot to be the least legible cloud, as with a word count of 150, many of the terms were small, and as such became illegible as the cloud was shrunk to fit the screen size. By restricting the word count to 100, smaller terms were eradicated.

The second step was to edit the size of the clouds differently. Wordle JPEG files were resized by picture editing software to the correct proportions before being added to the HTML. TagCrowd offered a HTML version of the cloud, so this was used and display size controlled by CSS. This provided clearer clouds, with no unreadable terms.

Thirdly, the clouds were made the same width as the screen. Initially, these were in line with the rest of the text, much as a narrative summary would be. Even by making the JPEG files clearer, some of the words remain too small to be readable when taking this form. Making the clouds the same width as the screen allowed participants to read every word, if necessary.

Various issues with the questionnaires were shown up by the pilot. Many of the questions referred specifically to the clouds the participant had just seen, however as they had seen a total of 6 clouds, asking the participant to remember specific details was overloading. To alleviate this problem, sample clouds shown in the questionnaires were changed to the clouds produced for the individual participant.

As participants would now see the summary clouds more than once, and while answering questions asking them to reflect on certain properties, measuring the time spent on each summary became redundant, and as such was removed.

Smaller changes to questionnaires involved re-wording some unclear questions and changing vocabulary. It was found that participants found it hard to indicate the position they had reached with writing their dissertation, as so much of the work was ongoing. This was changed so participants could give an indicative percentage of their dissertation progress. Where the terms 'edited' and 'unedited cloud' had been used, these were changed to 'small cloud' and 'large cloud' respectively, as the participants were unaware of the editing taking place for each cloud.

3.6 Limitations of research

There are several limitations to this evaluation. Only six participants were used, which limits the conclusions of the research. Although three word clouds were utilised, only two different types were represented; a word cloud and a tag cloud. These differ in the way they present information; although both represent word frequency by text size, word clouds present these in a random pattern, where tag clouds present lines of words in alphabetical order. Other clouds identified in the literature review such as tree clouds, phrase nets and various ManyEyes applications present words in additional ways, and may produce different results, however use of these is beyond the scope of this research, mainly due to copyright restrictions. ManyEyes is owned and run by IBM, and content for word clouds has to be uploaded, saved and be available for public display. Uploading the articles used for this research would break copyright legislation.

Wordle allows colouring of word clouds, including a black and white option. TagCrowd clouds are always variegated shades of blue. As mentioned previously, consideration was given to having a cloud just in black and white, given that summaries are more likely to be printed without colour. This was deemed to be outside the scope of this research, as it would involve using either a fifth cloud, which might have overloaded participants, or randomly selecting a cloud each time to produce in black and white. Randomly colouring various clouds may influence other selection criteria. To gauge a general attitude to black and white clouds, participants were shown a colour cloud from their original summaries, and its corresponding black and white copy. Only general impressions can be taken from these.

Three different articles were presented and their content assessed by various summaries. Each article was studied on an individual basis. At no point were the articles compared to each other, so word clouds are assessed by relevance to a single article, any assessment of usefulness will not apply to scrutiny and comparison of word clouds across various articles.

As the study focused on the task of article selection, participants were only asked to indicate whether they would go ahead and read the article, they did not actually read it as part of this research. It may have been the case that in the course of their detailed, post-research proposal reading they had read the article in question. This could affect article acceptance or rejection, however it was generally assumed that they had not read the article, only the usefulness of the summary in regard to selection was evaluated. As such, the results can only apply to the task of article selection; it cannot show that the summaries were good representations of the original article, although this was assumed.

Using each participants' individual dissertation topic prevented agreement between participants being measured. There is no baseline for comparison, as there is no control article and summary to compare participant responses to. Assessment was based on participant relevance assessment to their individual topic.

All indications of usefulness are based solely on the opinion of the participant. Although this was specifically chosen as they are assumed to be most aware of their information needs, it is not an infallible measure as it is based on preference. It only provided an indication of what participants found useful for this particular task and at a particular stage of their dissertations.

3.7 Ethics

As this research involves working with human participants, several ethical issues were addressed. These included the recording of participant responses, confidentiality of participants and ensuring data was stored securely both during and after the research had been completed. These issues were primarily addressed by keeping to university procedures and ethical guidelines, and by completing an ethical clearance form. By keeping to these guidelines, the specific issues identified were addressed by gaining informed consent from all participants, allowing withdrawal from the research at any time, ensuring non-disclosure of participants details by using letters instead of names, and only allowing my supervisor and myself to have access to any information collected.

Chapter 4: Findings and Discussion

4.1 Results

The evaluation was carried out in one-to-one sessions, as per the methodology (see Figure 1). Once the evaluation had been piloted, and appropriate changes made, six participants submitted their topics and literature reviews, and then viewed word cloud summaries which represented articles specific to their individual dissertation research.

Questionnaires were completed before and after viewing summaries for one article and again after viewing all summaries, and comments were then recorded regarding opinions on colour vs. black and white clouds (see Appendix C).

The sessions with the participants varied in length from half an hour to an hour and a half. Participants spent varying amounts of time considering the clouds, reflected in the wide variation of how much of each cloud each participant read.

4.1.1 'Before' questionnaire

Participants' initial impressions of summaries were gauged in the first questionnaire, to find out what they felt was needed in a "good" summary. These showed both what information was expected in a summary, and how important layout was to evaluation.

Each of the six participants rated each possible component on a scale of 1 (not important) to 5 (very important), giving each component a possible score out of 30. As such, the higher the score, the more important the component was overall to the participants (see Table 2).

All participants found summaries useful when searching for articles. That they covered main points, included key words and key phrases from the original article rated highly with all participants. Key words were considered more important than key phrases, although other measures of context, such as that the summaries outlined arguments, gave background information to a topic, and explained context, were all rated highly. One participant added that giving details of any methodology was very important to be included as part of a summary.

Table 2: Scores for importance of specific summary components

30	Cover main points of an article
28	Key words
27	New information
26	Outlines arguments
25	Explains context; includes background information; short/can assess quickly
24	Key phrases
17	Content follows same order as article
16	Includes visual information
4	Methodology (one participant only)
30	Maximum score

Other information expected in a summary included novel details in the form of new information, such as arguments and information not previously seen in relation to the topic, with visual information less necessary, however two participants rated visual information at 4, which effectively doubled this score.

Layout of summaries was also rated. That summaries were short and easy to assess was important, and several comments were made about the general usefulness of including summaries referred to time-saving, length, and their contribution to reducing information overload. That the content followed the same order as the article was much less important.

Other comments made about summaries included their usefulness for filtering a large number of possible articles, and that they streamlined the literature review process.

Overall, participants placed quite high demands on summaries, and the demands they make for context, time saving and relevant terminology are easily met by traditional narrative summaries. Word clouds should also be able to meet the majority of these demands, depending on whether the gisting context given by the cloud is enough for participants to determine arguments and background information.

Participants were looking for articles that matched their research topic and key words, and articles that offered new, previously unseen information. For summaries to perform and be of use for this task, responses show that summaries need to be a

clear representation of the main points and arguments of the article, and that these are represented in a way that is short and quick to access.

4.1.1.1 *Dissertation stage reached at time of study*

The stage each participant had reached with their dissertation was ascertained, as it was a concern that the further along a participant was with their research, the more specific their information needs would be. No participant considered themselves further than 50% through their research, suggesting that all were still interested in finding new information and actively searching for articles. This is reflected in the rating given to new information being included in summaries, and indicates that the filtering process of the summaries would still be relevant to their information needs.

4.1.2 'During' and 'after' questionnaires

All of the summaries encouraged participants to read the article, except one. The large cloud (an unedited Wordle) was most likely to make participants read the article, and was the cloud that was liked best. The tag cloud (an edited Tagcrowd) was second on both counts, but scored low, followed by the small cloud (an edited Wordle) (see Figure 4).

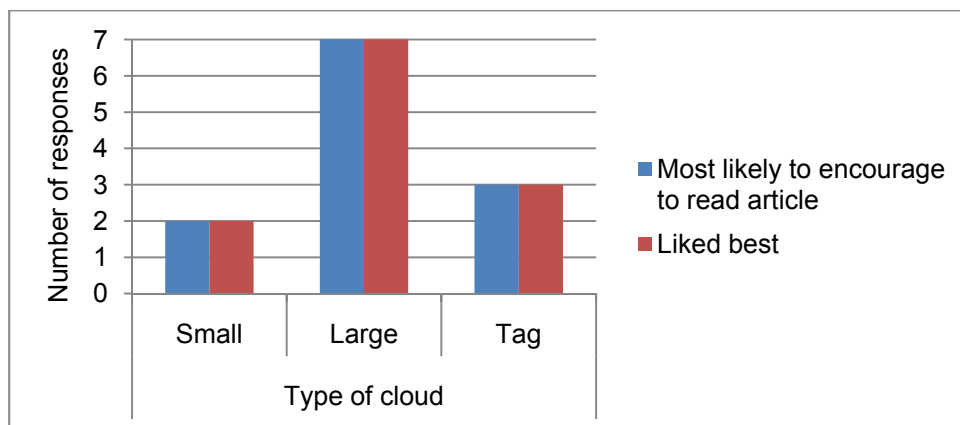


Figure 4: Number of responses for clouds most likely to encourage participants to read the corresponding article, and which cloud type was liked best

All types of cloud allowed participants to accept or reject the article confidently. Each participant was asked to indicate as a percentage how much of each cloud they felt they had read. These responses showed that the decisions whether to accept or reject an article were based on reading less than 100% of each cloud (see Figure 5), some participants read considerably less than others. Overall, there was very little difference between how much of each type of summary was read, however the size of each cloud was relative so reading less of the large cloud may still equate to the same number of words.

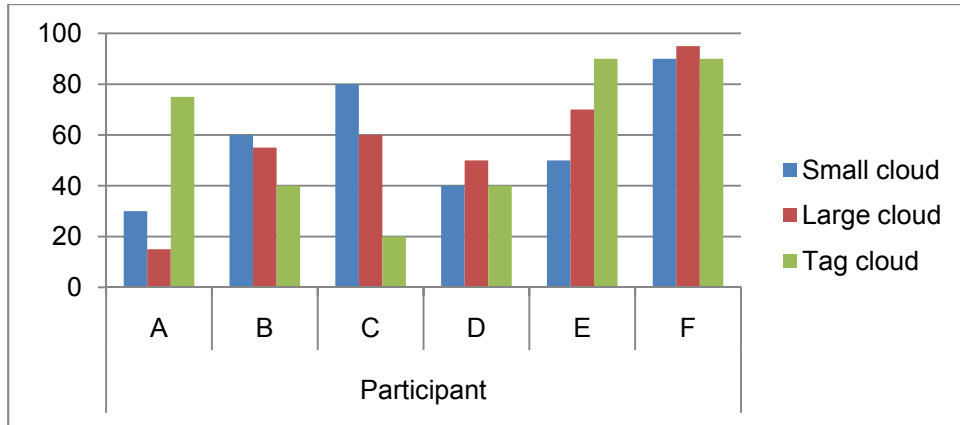


Figure 5: Percentage of each cloud participants said they had read

Participants were asked to comment, using a percentage scale, on their general impression of the content of the article, and their understanding of the articles positioning and argument from the clouds. Again, very little difference in impression existed between summaries. With a slight margin, large clouds were rated as giving the best general impression and understanding of argument (see Figures 6 and 7).

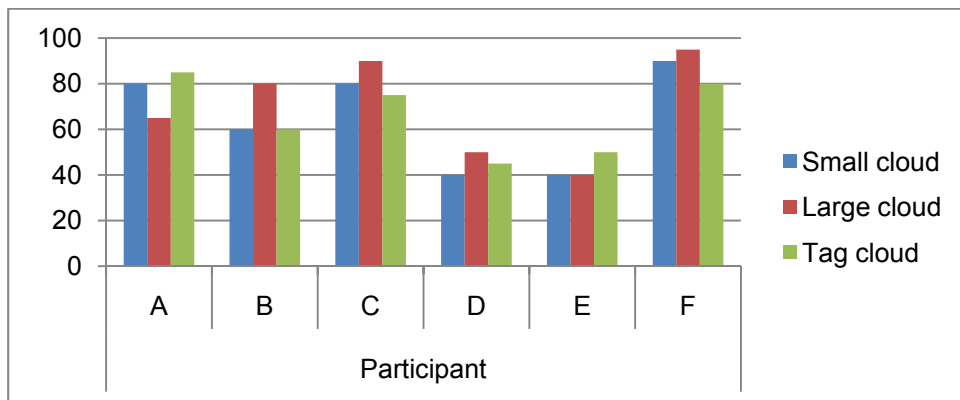


Figure 6: Percentage assessment of general impression of article content

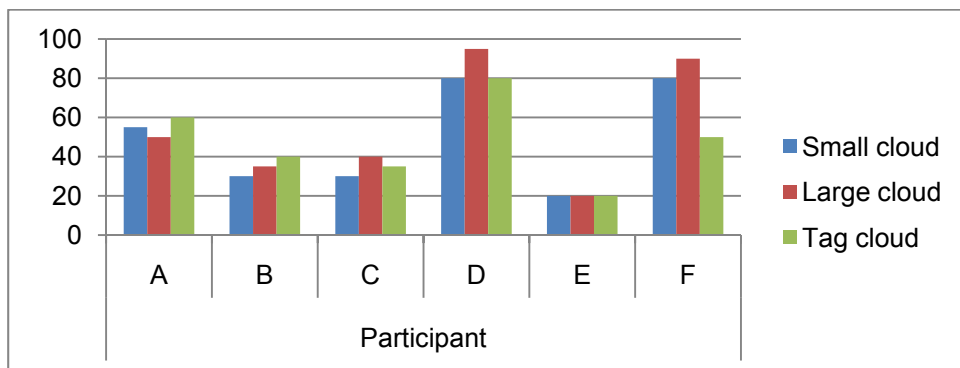


Figure 7: Percentage assessment of understanding of position and argument of the article

4.1.3 General cloud comments

Participants were asked what they found useful about each specific type of summary. Various responses were typical to all summaries; that summaries featured large, relevant words and key words, and that they were succinct, concise and quick to read.

Unhelpful features again attracted typical responses; that there were fewer specific, unhelpful words, potentially unhelpful words were emphasised, anti-key terms were shown, as well as various aesthetic reasons.

When asked what specifics of the summaries encouraged participants to read the article, responses included; key words, phrases and relevant terms being included, words suggesting desirable arguments and the, 'cumulative effect of numerous relevant terms'.

Generally, the features that participants found useful across all summaries were the same as those identified in their initial expectations, emphasised again in the final section of the fourth questionnaire, where the prominence of key words and that summaries covered key topics were judged the most relevant contributing aspects to choosing articles. "Unhelpful features" are the antithesis of these recurring summary features.

4.1.4 Specific cloud comments

More specific features were raised regarding certain types of summaries. These specific features were mentioned by only one or two participants, in addition to the themes that emerged for all variations of the clouds.

Small clouds were labelled useful as they offered context; unhelpful as there were no cited authors, findings, position or methodology, and due to the small amount of content, there was a fear that significant information could be missing. Ironically small clouds, as mentioned in the methodology, had been edited in such a way to filter out the 'background noise' of assumed irrelevant or redundant material from the article. That participants were interested in cited authors (Figure 8) – not something that would usually be included in a traditional narrative summary, suggests that the overall gisting process is coming from a few words, which, although not necessarily key words, when combined suggested a bigger picture. A selection of cited authors or words from a methodology (probably missing from the smaller cloud as the methodology would be a relatively small section, possibly



Figure 9: Edited word and tag cloud including the terms 'pre-prints' and 'post-prints'

Several participants offered general comments on how useful they had found the summaries. Two suggested that they were time saving, and two that they would still want to use an abstract alongside such summaries to be certain they fit the topic. One participant commented that Wordle gave more detail and clearer words, and that overall summaries were an easy breakdown of topics and words for an overview.

4.1.5 Colour vs black & white

Participants were shown the same word cloud twice, one in colour and one in black and white. Most participants preferred the colour version, giving reasons that it made details clearer, helped to pick out words and alerted you, that it was attractive, and that you wanted to read everything.

Comments regarding the black and white cloud included that it was easier to read all the words when in black and white, but also that they were similar to writing everything in capital letters, in that it was not so easy to read and encouraged skimming.

That participants showed a preference for the colour copy of the cloud may affect their use as summaries, as summaries are often printed in black and white. This was only tested using Wordle clouds, so the same may not apply to a tag cloud layout.

4.2 Discussion

All participants asked for copies of the articles after viewing the summaries, showing that they felt the impression they got from the word clouds was adequate to judge the articles as relevant, and also that they had not come across the articles during their own searching, so knowledge of the full text did not influence the results. As participants would go ahead to read the articles, word clouds were useful in this situation of selecting and filtering search results and clouds satisfied the participant's information needs.

Large clouds were found most useful in terms of article choice, gisting and for feel of an article's argument, and were clearly the most useful choice for all participants. Compared to what was first expected in summaries (see Table 2), large unedited clouds arguably offer more than just the main points of an article, if the top 50 words are considered the main points. The large range of words however, was cited as one of the reasons for preferring the large cloud. Participants seem to expect more content in a cloud than from a narrative summary.

One of the issues brought against narrative summaries in the literature review was that an assumption exists that the objective meaning of a text can be extracted, and that this assumption denies reader interpretation (Hirst 2007; Morris 2010). Where a narrative summary gives the main points of an article in sentences which need reading and sense-making is already enforced, clouds visually represent main points (large words) without enforced sense-making, and participants may prefer the larger cloud as it gives more terms to draw their own conclusions and make their own sense of an article. Background and context rated highly in initial expectations of summary, and with clouds, are participant-determined. Cloud summaries offer an actual objective view of the text, representing word counts only, and allow for reader interpretation, answering to the concerns raised in the literature review (Hirst 2007; Morris 2010).

In the case of clouds, participants cited that the cumulative effect of words gave an impression of argument. Large clouds possibly offer a greater effect by displaying more words.

The literature review initially led to the edited clouds being produced, due to the concerns of too many words being less intuitive (Yang, et al 2003), and the broad focus of the articles on layout and presentation of the clouds (Hearst and Rosner 2008; Gambette and Véronis 2009; Cui, et al 2010). By editing the clouds, both by reducing words displayed and through amalgamating synonyms, plurals and key

phrases, these concerns were addressed as some of the context of the original article had been preserved.

The results show however, that this was not preferred by participants in this study. Rather, the benefits found from using clouds in catalogue environments were applicable to the large, unedited clouds, rather than the smaller, edited versions. Benefits such as revealing controlled vocabularies not initially evident to the user are reflected in participant responses such as the discovery of new information. Distance between terms, rather than confusing or hindering participants seems to have encouraged discovery, as suggested by Hearst and Rosner (2008).

Edited clouds address some of the usability and distribution issues raised by the literature, however in preferring large unedited clouds, participants seem to prefer to have control of the discovery, and not to have context guided by key phrases and reduced word display. This was perhaps initially suggested by the first questionnaire, when key phrases rated significantly lower than key words, and in criticisms of traditional summary evaluation, where user interpretation was denied (Hirst 2007; Morris 2010).

Edited clouds, although not the overall preferred choice, were useful to some participants. One commented specifically on phrases being joined together as useful to work out how it fit with their specific research, and two other participants recorded phrases as specific reasons that would make them want to read the articles (the words/phrases were: Myers Brigg Type Indicator, post-prints, pre-prints – see Figures 10 and 9 respectively). These were not key words or phrases for their research, however the mention of them was suggestive of discussion relevant to their topic.

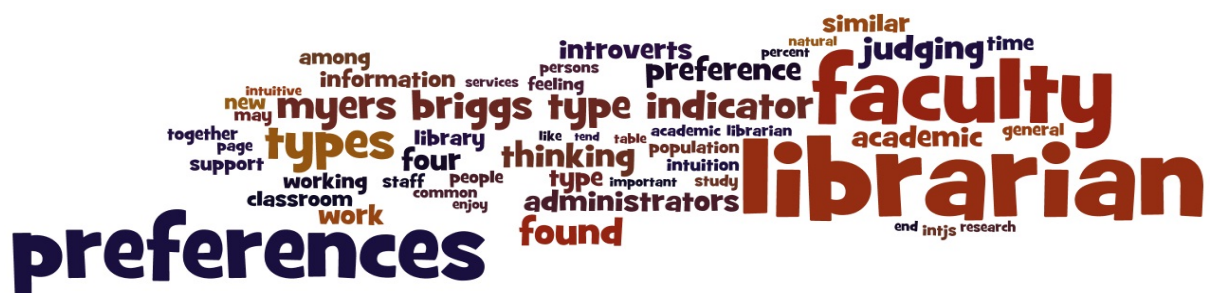


Figure 10: Word cloud showing phrase 'Myers Briggs Type Indicator'

ManyEyes, an IBM cloud generator not used in this study due to its copyright restrictions, generates clouds that represent two word phrases in a similar way to the editing done here. Although participants preferred larger clouds, the majority of

comments justifying this decision refer to the broad coverage and as such word count of the cloud, if edited clouds were produced to the same word count used for large clouds, results may be different.

Most of the comments revolved around the content (or lack of) of the clouds, very little was said about the different layout and colour of the tag cloud, which produced blue text in alphabetical order. One participant found that some of the words in the tag cloud were not clear as they could overlap, and another that it was more organised. The font of the cloud had been edited with HTML to prevent aesthetic impressions impacting on participant consideration of the content, however again it is interesting in relation to the concerns raised in the literature review over word distribution (Cui, et al 2010) that participants did not as a whole recognise the alphabetisation or find that aesthetics affected their assessment of the clouds. It may be that they did not recognise that aesthetic reasons had this effect, as previously discussed, one participant did mention noticing phrases in the tag cloud “not featured” in other clouds, although the tag cloud content exactly mirrored that of the small cloud.

The literature review revealed that word clouds had been used to scrutinise and compare texts to gain insight (Gill and Griffin 2010; Schwenkler 2008). There is some suggestion of participants having done this to get a gist of the argument and position. One participant spotted the terms ‘guilt’ and ‘bond’ in clouds (see Figure 11), and suggested that, although not major terms, these suggested a desirable argument. Interestingly, these were only picked out in the Wordle generated clouds, and the participant estimated reading less than 50% of the tag cloud. Rather than just using the key words present in the text to determine relevance, a deeper insight into the argument the text presented was revealed.





Figure 11: Word clouds where the words 'guilt' and 'bond' were noted, suggesting desirable arguments

That another participant saw phrases in the tag cloud but not in its corresponding small Wordle, with identical text, the issue of clouds being independently sized and relational (Cidell 2010) becomes a possible issue for scrutiny. This participant owned to reading 90% of both the small and tag clouds.

How much of each cloud is being read, and possibly what arguments and context participants were specifically looking for are of interest when looking at how far clouds were scrutinised. Ultimately, participants found summaries useful and wanted to read the articles, their individual enforcement of context onto clouds evidently revealed different aspects of the text from each cloud, with the large cloud being the cloud making them most confident of their own analysis and interpretation.

One of the participants, who speaks English as a second language, remarked In the feedback that cloud summaries were quicker to evaluate than 'normal' summaries. This is possibly because this particular participant found translation of single words quicker than sentences.

Participants were each shown a careful selection of analytical and descriptive articles, as the literature review suggested that cloud summaries may better represent descriptive content (Kuo, et al 2007). Considerable time was taken in preparation for sessions with the participants to ensure each was presented with a range of both descriptive and analytical articles to prevent this influencing results. There was no significant difference between the two types of articles here, although the task differed slightly. Kuo's study focused on participants answering questions using either a word cloud or a "normal" search engine, and it was these questions that participants took more time to answer from a cloud. Gisting and overall heuristic impression of the cloud was studied here, and there was no significant difference between the number of analytical or descriptive articles chosen to read.

Chapter 5: Conclusion

That participants were prepared to go ahead and read the articles shows that the clouds answered an information need and were useful in this situation of article selection. Word clouds met many of the initial criteria expected of summaries, and characteristics such as key words and covering main points were specifically mentioned as useful and contributing to the decision to read the article. The implications of this are considered in relation to the original objectives (see Section 1.3).

5.1 Range of current summaries

A range of summaries and their uses were identified by the literature review and through the evaluation. Human or computer, indicative or informative summaries (Mani, et al 2002) were identified as being used for document retrieval, skimming overviews and browsing. Participants in the study echoed these demands by prioritising short, easy to read summaries which saved time and helped to combat information overload. The literature review identified an increasing range of summary types to answer technological demands such as screen size issues, and also utilisation of Web 2.0 technologies for novel summary production (Gambette & Véronis 2009; Yang, et al 2003). Word clouds had not been utilised specifically as summaries for document selection, the main use found for them in the literature was to aid traditional catalogue searching (Mayfield, et al 2008; Olson 2007; Seifert, et al 2008).

5.2 Current issues in the production of summaries

Although production of summaries is resource intensive, with a lot of time and skill put into their creation, it would be unusual to find an academic article without an accompanying summary or abstract. Through the literature review, current issues in evaluation of these summaries were analysed, with the main issue being that interpretations of articles given in narrative summaries were found to be subjective (Morris 2010; Johnson 1995; Hirst 2007). Macro and micro information (Zhan, et al 2009) were identified as a possible way to reduce this subjectivity, but this heavily relied on low word counts.

In the study, an attempt was made to avoid these issues by utilising a heuristic method to assess usefulness of the summary in the frame of a participant's individual information needs. Although the cloud summaries would not have met the criteria for "gold standard" objectives, as separating words and terms prevents them from being linguistically well-formed, participants owned to finding clouds useful,

and justified this with explanations as to specific features that were “good”. The subjectivity of individual interpretations was allowed and encouraged in this study, allowing multiple interpretations. This evaluation was successful in gauging fulfilment of individual need, rather than the needs of pre-determined tasks.

Although clouds were useful to individuals and their individual research topics, as no continuity or control was introduced to the topics, no comparison of usefulness can be made between different subject areas. Pre-determined tasks set clear objectives such as categorisation that were not appropriate to be replicated here. The responses to these tasks could be used as an objective measure of cloud performance, and by using heuristic evaluation, there is a possibility that although the clouds were useful to these participants, their individual topics happened to be represented well by word clouds.

5.3 Use of Web 2.0 as indicator of content summaries

Word clouds were identified in the literature review as the main Web 2.0 tool utilised for summary generation. These were confirmed as being used mainly for gisting (Rivandeniera, et al 2007). As indicator of content summaries, various issues were raised with utilising word clouds as summaries.

Concerns have been raised over the distribution and number of words being misleading (Yang, et al 2003) and over their relational size (Cidell 2010; Hearst & Rosner 2008). The evaluation has shown that for the scope of this study, these issues were not relevant, participants actually preferring more random distribution and more words (demonstrated through the preference for large clouds). Relational size was not assessed here, as clouds for separate articles were not compared.

A concern raised in the literature review was that analytical articles would be less well represented by word cloud formats than by descriptive articles (Kuo, et al 2007). Again, evaluation within the scope of this study has shown this was not an issue, both types of article were represented and the results were consistent in the case of the large cloud. The large cloud was considered a good indicator of content summary for the majority of articles.

It was suggested that lists of words in frequency order performed better than their corresponding clouds (Rivandeniera, et al 2007), however, as this assessment was made from a study of accuracy, it again falls outside of heuristic evaluation.

5.4 Appropriate situations

Indicators of content were shown by the literature review to be mainly used in library catalogue search situations (Mayfield, et al 2008; Olson 2007; Seifert, et al 2008), and were found appropriate for use in this way. As such, the evaluation also replicated a search environment. Reasons why word clouds may not be appropriate in certain situations were raised above (see Section 2.3.3) while assessing the uses of Web 2.0 summaries.

5.5 'Good' summaries

A full discussion of the features participants used to gauge how "good" a summary is was covered in the discussion section (see Section 4.1.1). Participants had a shared set of assumptions, strongly related to the characteristics of narrative summaries (covers main points, outlines argument, explain context, short).

5.6 Evaluate usefulness of summaries

Large clouds were the most useful and preferred, and some of the reasons these were preferred and other clouds found less useful were for additional demands, not initially listed as expected in summaries. Large clouds were preferred for offering more details than simply the main points, however this may have been expected as more words for participants equated to more context, a factor initially considered important.

That cited authors were singled out as useful features in large clouds, although these are not featured in traditional summaries suggests, as previously discussed, that participants prefer to bring their own experience and context to the summary (Morris 2010). Participants bring their own conceptual framework to the cloud, and by selecting large clouds and various details in them, this indicates that individual interpretation is a useful tool when selecting articles.

Small word clouds and tag clouds were similarly rated, however tag clouds received much fewer comments and feedback from participants. Comments that were recorded included that they were more legible and that key terms were seen that were not found in other clouds, which implies a higher rating than given in formal feedback.

Tag clouds may have been found less useful due to aesthetic reasons. Although an attempt was made to minimise the impact of their layout, all tag clouds were shown in various shades of blue. Taking the comments about black and white clouds into

consideration, it may be that these were found less useful for similar reasons; that they were not as eye-catching or engaging as the multi-coloured Wordles.

Clouds generated by Wordle also had a novelty factor that participants may have found lacking in the tag clouds. Participants may be more familiar with tag clouds through use of sites such as Delicious¹⁶ and Flickr¹⁷, where Wordle clouds are less widely used. Wordle generated clouds use a wide range of colours, fonts and shapes which, in contrast may have made the straight lines and regular rectangle shape of the tag cloud less appealing and engaging.

Even taking novelty into consideration, small clouds did not perform very differently to tag clouds when it came to article selection. Participants preferred the 100 words given in the larger cloud, although the literature review suggested that phrases would have been more useful and less distracting for gauging context (Yang, et al 2003).

The layout of the final questionnaire possibly affected feedback for tag clouds. Clouds were presented in a random order when viewed as search results, however, the questionnaire did not reflect this randomisation, asking questions about large, small and then tag clouds (see Appendix C). As these questions were quite intensive for participants and demanded a lot of information, feedback to tag clouds may have suffered, for being on the last page, and for asking an identical question which called for original answers.

Overall, clouds were useful in an unanticipated way, in that participants found a degree of abstraction in the dissociated terms which allowed them to draw conclusions and analyse the text further than a general overview. The cumulative effect of certain terms in the clouds gave an impression of argument and direction of the article. This was suggested in the literature review where clouds had been used to analyse the US 2008 election (Schwenkler 2008), and GMC documents (Gill and Griffin 2010). That participants were scrutinising and analysing clouds in this way may explain why cited authors were found useful, participants could draw more informed conclusions with the knowledge of similar articles.

The evaluation did not show whether participants were looking for a specific argument or theme when scrutinising clouds. As all participants were less than 50% through their dissertation, it is implied that further information was still of use to

¹⁶ Delicious <<http://delicious.com/>>

¹⁷ Flickr <<http://www.flickr.com/>>

them, however as everyone had started their research, it may be necessary to have these reference points (such as authors) in order to scrutinise the disassociated terms.

Coloured clouds were found more useful than black and white clouds which has implications for use. Although coloured clouds would be viable in an online environment, and may even be appropriate for small screens, the charges involved in printing in colour would increase publication costs. This would also be an issue for various colour blind users.

5.7 Limitations

The limitations of the study have been discussed in the methodology (see Section 3.6). That the evaluation took place with six participants limits the scope of the conclusions, as some rely on the comments of only one or two participants.

The possible fatigue of participants needs to be taken into account as a large number of clouds were presented to participants and the final questionnaire (see Appendix C) was quite demanding in that lengthy responses were required.

Only the layout of a tag cloud was used, the functionality of clicking a term which then takes you through to the word in the document was not offered by TagCrowd.

5.8 Recommendations for further research

Various issues were raised in the study which could be investigated through further research.

- This study was relatively small, using six participants and their individual topics. Conclusions that have been drawn could be tested further using a larger number of participants and research topics. More careful study could be made of the differences in representation of relational and descriptive articles. This would need to involve more specific questions regarding the overall feel participants felt they gained of the article.

Recruiting further participants for the same task should be feasible, as many postgraduate students produce research proposals and carry out dissertations every year. Dependent on the size of the study, a lot of resources would need to be allocated to searching for appropriate articles and setting up the clouds, as this was the most intensive activity in the methodology.

- As it is not entirely clear if the large clouds were preferred for the context they offered or simply because of the word count, further research could present both edited and unedited clouds with word counts of 100. This should more clearly distinguish whether participants prefer enforced context from the phrases and editing done in the edited clouds, or context they have taken themselves from unedited text. This could still initially be a small study, to see if the same preference for large clouds exists.

Although this study would be feasible, there is a possibility that participants may be confused and unable to distinguish the clouds. Especially due to colours in Wordle, words may seem as if they have been moved around rather than joined together to provide context.

- One participant in this study spoke English as a second language and found clouds easier to use than narrative summaries. It was inferred that this may be because words are easier to translate than tense-ridden sentences. Further research could focus specifically on students who speak English as a second language and cloud use. Research could involve reading both a narrative summary and a cloud summary, and usefulness gauged in the scope of the individuals feeling of comprehension. A possible issue that may arise is communication issues in feedback.

5.9 Concluding remarks

Cloud summaries have been shown to be useful when selecting articles from a search. They encouraged participants to read the corresponding article by meeting the criteria participants set for summaries.

Large clouds were preferred, and further research needs to take place in order to distinguish whether this was because of their word count, or if users prefer to determine their own opinion and view of context.

Through scrutiny and analysis of clouds, a degree of abstraction was obtained that enabled certain participants to interpret the article that the cloud represented. The ability to form a personal opinion of article content is denied in narrative summaries, as an interpretation of the article is already presented. Word clouds, as indicator of content summaries, answer one of the main criticisms of narrative summaries in allowing individual interpretation. This suggests that word clouds are appropriate for use as summaries for article selection.

Bibliography

Alfonseca, E. & Rodriguez, P., 2003. *Description of the UAM system for generating very short summaries at DUC-2003*. < <http://alfonseca.org/pubs/2003-duc.pdf>>, [accessed 16.06.10].

Burmeister, T., *Enthusiasm for word clouds – musings on indexing*. <<http://tburmeister.wordpress.com/2010/04/02/enthusiasm-for-word-clouds-musings-on-indexing/>>, [accessed 12.08.10].

Cidell, J., 2010. Content clouds as exploratory qualitative data analysis. *Area*, In Press.

Cooper, C. A., Collins, T. A. & Knotts, H. G., 2009. Picturing political science. *PS: Political Science and Politics*, 42(2), 365-365.

Cosh, K. J., Burns, R. & Daniel, T., 2008. Content clouds : classifying content in Web 2.0. *Library Review*, 57(9), 722-729.

Cui, W., et al., 2010. Context preserving dynamic word cloud visualisation. *IEEE Pacific Visualisation Symposium 2010*. <http://ieeexplore.ieee.org/xpl/freeabs_all.jsp?arnumber=5429600>, [accessed 01.02.10].

Dalianis, H., et al., 2003. *From SweSum to ScandSum - automatic text summarisation for the Scandanavian languages*. < <http://people.dsv.su.se/~hercules/scandsum/ScandSumArsbog2002.pdf>>, [accessed 08.06.10].

Freedman, T., 2010. Word cloud shoot-out. *Computers in Classrooms* [online], 21.04.10. < <http://ymlp.com/z0W039>>, [accessed 12.06.10].

Gambette, P. & Véronis, J., 2009. Visualising a text with a tree cloud. *IFCS'09 International Federation of Classification Societies Conference*. <<http://hal-lirmm.ccsd.cnrs.fr/lirmm-00373643/>>, [accessed 09.06.10]

Gibson, O., 2010. *Tiger Woods begs forgiveness for 'selfish and foolish' behaviour*. <<http://www.guardian.co.uk/world/2010/feb/19/tiger-woods-begs-for-forgiveness>>, [accessed 12.08.10].

Gill, D. & Griffin, A., 2010. Good Medical Practice: what are we trying to say? Textual analysis using tag clouds. *Medical Education*, 44(3), 316-322.

Giora, R., 1985. A text-based analysis of non-narrative texts. *Theoretical Linguistics*, 12(2-3), 115-135.

Gottron, T., 2009. Document word clouds: visualising web documents as tag clouds to aid users in relevance decisions. *Proceedings of the 13th European Conference on Research and Advanced Technology for Digital Libraries*.
<<http://portal.acm.org/citation.cfm?id=1812814>>, [accessed 03.07.10].

Harper, S. & Patel, N., 2005. Gist summaries for visually impaired surfers. *Proceedings of the 7th international ACM SIGACCESS conference on computers and accessibility*. <<http://portal.acm.org/citation.cfm?id=1090785.1090804>>, [accessed 25.01.10].

Hearst, M. A. & Rosner, D., 2008., Tag clouds: data analysis tool or social signaller? *Proceedings of the 41st Hawaii International Conference on System Sciences*.
<<http://www.computer.org/portal/web/csd/doi/10.1109/HICSS.2008.422>>, [accessed 03.07.10].

Hirst, G., 2007. Views of text meaning in computational linguistics: past, present and future. In: Dodig-Crnkovic, G. & Stuart, S., eds. *Computation, information and cognition – the nexus and the liminal*. Newcastle-Upon-Tyne: Cambridge Scholars Press, 267-276.

Johnson, A., 1995. Automatic abstracting research. *Library Review*, 44(8), 28-36.

Kaminuma, E., et al., 2009. DDBJ launches a new archive database with analytical tools for next-generation sequence data. *Nucleic Acids Research*, 38(1), 1-6.

Kuo, B. Y-L., et al., 2007. Tag clouds for summarising web search results. *Proceedings of the 16th international WWW conference (WWW2007)*.
<<http://portal.acm.org/citation.cfm?id=1242572.1242766>>, [accessed 29.01.10].

Lester, D. F. & Robinson, M., 2009. Visions of exploration. *Space Policy*, 25(4), 236-243.

Li, Q. & Chen, Y. P., 2010. Personalised text snippet extraction using statistical language models. *Pattern Recognition*, 43(1), 378-386.

Luhn, H. P., 1958. The automatic creation of literature abstracts. In: Mani, I. & Maybury, M. T., eds. 1999. *Advances in automatic text summarisation*, 2nd ed., 15-23.

- Mani, I., et al., 2002. SUMMAC: a text summarisation evaluation. *Natural Language Engineering*, 8(1), 43-68.
- Mayfield, I., et al., 2008. Next generation library catalogues: reviews of ELIN, WorldCat Local and Aquabrowser. *Serials*, 21(3), 228-230.
- McKie, S., 2007. Scriptcloud.com: Content clouds for screenplays. *2nd International Workshop on Semantic Media Adaptation and Personalisation*.
<<http://www.computer.org/portal/web/csdl/doi/10.1109/SMAP.2007.51>>, [accessed 12.08.10].
- Morris, J., 2010. Individual differences in the interpretation of text: implications for information science. *Journal of the American Society for Information Science and Technology*, 61(1), 141-149.
- Olson, T. A., 2007. Utility of a faceted catalogue for scholarly research. *Library Hi Tech*, 25(4), 550-561.
- Over, P., Dang, H. & Harman, D., 2007. DUC in context. *Information Processing and Management*, 43(6), 1506-1520.
- Rivadeneira, A. W., et al., 2007. Getting our head in the clouds: toward evaluation studies of tagclouds. *CHI 2007*.
<<http://portal.acm.org/citation.cfm?id=1240624.1240775>>, [accessed 16.06.10].
- Ruch, P., et al., 2007. Using argumentation to extract key sentences from biomedical abstracts. *International Journal of Medical Informatics*, 76(2-3), 195-200.
- Schwenkler, J., 2008. Portrait of the candidate as a pile of words. *Boston Globe*. <http://www.boston.com/bostonglobe/ideas/articles/2008/08/03/portrait_of_the_candidate_as_a_pile_of_words/>, [accessed 12.06.10].
- Seifert, C., et al., 2008. On the beauty and usability of tag clouds. *IEEE Information Visualisation 2008 12th international conference*.
<http://ieeexplore.ieee.org/xpl/freeabs_all.jsp?arnumber=4577920>, [accessed 16.06.10].
- Spärck Jones, K., 2007. Automatic summarising: the state of the art. *Information Processing and Management*, 43(6), 1449-1481.
- Strobel, H., et al., 2009. Document cards: a top trumps visualisation for documents. *IEEE Transactions on visualisation and computer graphics*, 15(6), 1145-1152.

- Sweeney, S., Crestani, F. & Losada, D. E., 2008. 'Show me more': incremental length summarisation using novelty detection. *Information Processing and Management*, 44(2), 663-686.
- Veel, K., 2009. *Narrative negotiations: information structures in literary fiction*. Göttingen: Vandenhoeck & Ruprecht.
- Viégas, F. B. & Wattenberg, M., 2008. Tag clouds and the case for vernacular vision. *Interactions*, 15(4), 49-52.
- Viégas, F. B., Wattenberg, M. & Feinberg, J., 2009. *Participatory visualisation with Wordle*. <
[http://domino.research.ibm.com/cambridge/research.nsf/58bac2a2a6b05a1285256b30005b3953/6878415e7fd0efa48525766d00715513/\\$FILE/TR%202009.05%20wordle_final2.pdf](http://domino.research.ibm.com/cambridge/research.nsf/58bac2a2a6b05a1285256b30005b3953/6878415e7fd0efa48525766d00715513/$FILE/TR%202009.05%20wordle_final2.pdf)>, [accessed 12.06.10].
- Wei, F., et al., 2009. Applying two level reinforcement ranking in query oriented multi-document summarisation. *Journal of the American Society for Information Science and Technology*, 60(10), 2119-2131.
- Yang, C. C., Chen, H. & Hong, K., 2003. Visualisation of large category map for internet browsing. *Decision Support Systems*, 35(1), 89-102.
- Yang, C. C. & Wang, F. L., 2008. Hierarchical summarisation of large documents. *Journal of the American Society for Information Science and Technology*, 59(6), 887-902.
- Zhan, J., Loh, H. T. & Liu, Y., 2009. On macro- and micro- level information in multiple documents and its influence on summarisation. *International Journal of Information Management*, 29(1), 57-66.

Other Sources Consulted

Alonso, M. I. & Fernández, L. M. M., 2010. Perspectives of studies on document abstracting: towards an integrated view of models and theoretical approaches. *Journal of Documentation*, 66(4), 563-584.

Assogba, Y. & Donath, J., 2009. Mycrocosm: Visual Microblogging. *42nd Hawaii International Conference on System Sciences, HICSS 09*.
<<http://www.computer.org/portal/web/csdl/doi/10.1109/HICSS.2009.836>>,
[accessed 12.08.10].

Berson, I. R. & Berson, M. J., 2009. Making sense of social studies with visualisation tools. *Social Education*, 73(3), 124-126.

Blythe, K., 2008. A faceted catalogue aids doctoral-level searchers. *Evidence Based Library and Information Practice*, 3(3), 76-79.

Brandow, R., Mitze, K. & Rau, L. F., 1995. Automatic condensation of electronic publications by sentence selection. *Information Processing and Management*, 31(5), 675-685.

Donoser, M., Wagner, S. & Bischof, H., 2010. Context information from search engines for document recognition. *Pattern Recognition Letters*, 31(8), 750-754.

Evelt, L. & Brown, D., 2005. Text formats and web design for visually impaired and dyslexic readers - clear text for all. *Interacting with Computers*, 17(4), 453-472.

Fung, C. C., Thanadechteemapat, W. & Wong, K. P., 2009. Iwise, and intelligent web interactive summarisation engine. *Proceedings of the 8th International Conference on Machine Learning and Cybernetics*.
<http://researchrepository.murdoch.edu.au/773/1/Published_Version.pdf>,
[accessed 03.07.10].

Furnham, A. & McClelland, A., 2010. Word frequency effects and intelligence testing. *Personality and Individual Differences*, 48(5), 544-546.

Giustini, D. & Wright, M-D., 2009. Twitter: an introduction to microblogging for health librarians. *Journal of the Canadian Health Libraries Association*, 30, 11-17.

Hjorland, B., 2007. Information: objective or subjective/situational? *Journal of the American Society for Information Science and Technology*, 58(10), 1448-1456.

- Lilly, E. B., 2001. Creating accessible websites: an introduction. *The Electronic Library*, 19(6), 397-404.
- Liu, F. & Liu, Y., 2010. Exploring correlation between ROUGE and human evaluation on meeting summaries. *IEEE Transactions on Audio Speech and Language Processing*, 18(1), 187-196.
- Liu, S., 2009. Experiences with and reflections on text summarisation tools. *International Journal of Computational Intelligence Systems*, 2(3), 202-218.
- Mani, I., 2001. *Automatic summarisation*. Philadelphia: John Benjamins.
- Moens, M-F. & Dumortier, J., 2000. Use of a text grammar for generating highlight abstracts of magazine articles. *Journal of Documentation*, 56(5), 520-539.
- Ou, S., Khoo, C. S-G. & Goh, D. H., 2008. Design and development of a concept-based multi-document summarisation system for research abstracts. *Journal of Information Science*, 34(3), 308-326.
- Passant, A., et al., 2008. Microblogging : a semantic and distributed approach. *Proceedings of the 4th workshop on scripting for the semantic web, CEUR workshop proceedings*. <<http://www.semanticscripting.org/SFSW2008/papers/11.pdf>>, [accessed 01/02/10].
- Pembe, F. C. & Gungor, T., 2009. Structure-preserving and query-biased document summarisation for web searching. *Online Information Review*, 33(4), 696-719.
- Philbrick, J. L., Cleveland, A. D. & Pan, X., 2009. *Positioning medical libraries in the world of web 2.0 technologies*. <http://espace.library.uq.edu.au/eserv/UQ:179801/Web2.0Technologies_PhilbrickClevelandPan.pdf>, [accessed 12.06.10].
- Qin, J. & Yung, N. H. C., 2010. Scene categorisation via contextual visual words. *Pattern Recognition*, 43(5), 1874-1888.
- Rahman, A.F.R., et al., 2001. Automatic summarization of Web content to smaller display devices. *Sixth International Conference on Document Analysis and Recognition*. <<http://www.computer.org/portal/web/csdl/doi/10.1109/ICDAR.2001.953949>>, [accessed 25.01.10].

Teufel, S. & Moens, M., 2002. Summarising scientific articles: Experiments with relevance and rhetorical status. *Computational Linguistics*, 28(4), 409-445.

Appendices

Appendix A - Participant recruitment

Email text for participant recruitment

Hello,

I am a postgraduate student in the Department of Information Science. As part of my dissertation I'm doing research into what people find useful about document summaries. In order to look into this, I need some volunteers.

Volunteering would involve telling me about your current dissertation topic (either in person or by email), and also letting me have a copy of the bibliography from your research proposal. I'd use these to find articles and create some summaries. You would then need to come to a session on campus where you would look at and assess some summaries, completing questionnaires on what you found useful.

It is expected that in total this should take no longer than an hour and a quarter; a quarter of an hour to explain your topic and provide a bibliography (either in person or by email), and a maximum of an hour for the on campus session. I understand everyone is really busy with their own dissertations, and so the sessions will be really flexible with dates and times.

I would keep your topic, bibliography and questionnaire answers confidential, and you would be welcome to any of the articles I found in relation to your topic. I'm only going to use them for creating summaries, not any other purpose.

If you would like to take part, please send me an email and I'll get in touch to find out dates and times you might be available.

Many thanks,

Andrea

Appendix B – Participant session

Example of the screens a participant followed in a session

Loughborough University Library

Results

3 results found for: **INSTITUTIONAL REPOSITORIES, ZIMBABWE, OPEN ACCESS** in LISA: Library and Information Science Abstracts

Published Works 3

All Publication Types 3 Journals Peer-Reviewed Journals

Record # Go

<input type="checkbox"/>	1. Transforming access to research literature for developing countries Kirsop, Barbara; Chan, Leslie <i>Seriale Review</i> , vol. 31, no. 4, pp. 246-255, Dec 2005 View Record	Database: LISA: Library and Information Science Abstracts Descriptors: Research Developing countries Open access
<input type="checkbox"/>	2. Preserving African digital resources: is there a role for repository libraries? Lor, Peter Johan <i>Library Management</i> , vol. 26, no. 1-2, pp. 63-72, Dec 2005 View Record	Database: LISA: Library and Information Science Abstracts Descriptors: Africa Repository libraries Digital resources
<input type="checkbox"/>	3. Open access initiatives adoption by Nigerian academics Utulu, Samuel; Bolanihwa, Omolara <i>Library Review</i> , vol. 58, no. 9, pp. 660-669, Dec 2009 View Record	Database: LISA: Library and Information Science Abstracts Descriptors: Open access Nigeria Research

Show Short format Results per page: 10

© 2010 All rights reserved. | Privacy Policy | Terms and Conditions of Use | Contact Us Interface English

Participants choose one article

Loughborough University Library

Record View

Database LISA: Library and Information Science Abstracts

Title Transforming access to research literature for developing countries

Author Kirsop, Barbara; Chan, Leslie

Email Address b.kirsop@hope.ac.uk

Source *Seriale Review*, vol. 31, no. 4, pp. 246-255, Dec 2005

ISSN 0268-4012

Abstract

access archives authors available cost countries developing-countries development free health http
impact information initiatives institutions international journal libraries number
open-access org organizations provide publications publisher
research scientific users world www

Language English

Publication Year 2005

Publication Type Journal Article

[Next >](#)

© 2010 All rights reserved. | Privacy Policy | Terms and Conditions of Use | Contact Us Interface English

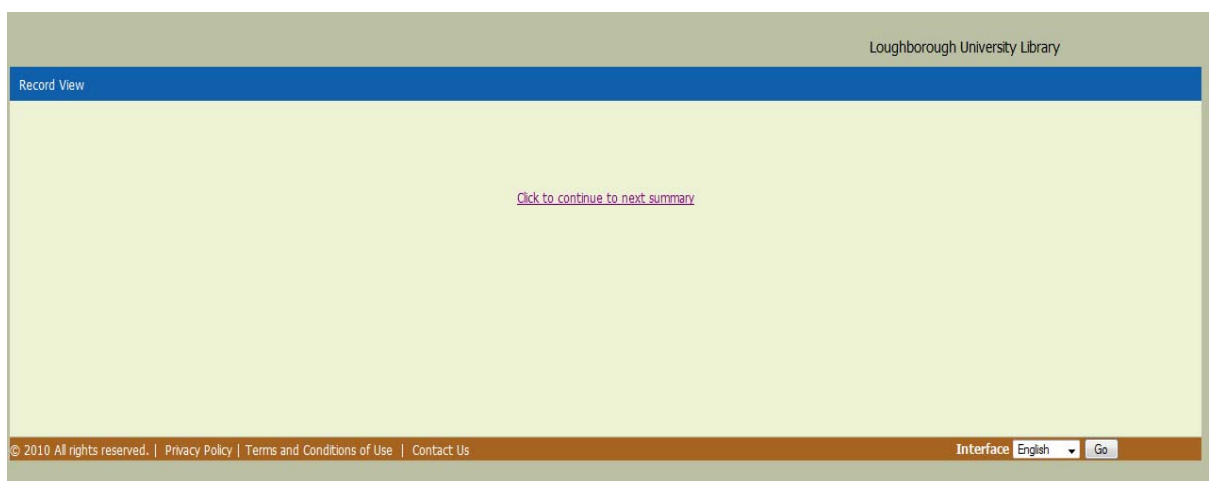
The first summary appears on screen, and when he/she finishes reading the summary, the next button is selected



A blank screen is shown, and participants click to move to the next summary



The second summary is shown, and the next button selected



A blank screen is shown, and again participants click to the next summary

Appendix C – Questionnaires

Examples of questionnaires

“Before” questionnaire

This questionnaire is to be completed before you view the summaries for the task. It is intended to gauge your impressions and assumptions of summaries.

1. Do you find summaries useful when you search for articles?

Yes No

a) If no, please provide reasons

2. How important do you think the following criteria are for a summary?

Not important----->

Very important

a) Short/can assess quickly	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
b) Cover main points of an article	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
c) Content follows same order as the article	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
d) Includes visual information	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>

3. How important is the following information as part of a summary?

Not important----->

Very important

a) Inclusion of key words	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
b) Inclusion of key phrases	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
c) Background information to topic	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
d) Outlines arguments of article	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
e) Includes new information	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
f) Explains context	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>

g) Other information you believe may be important to include in a summary (please specify)_____

	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
_____	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>

4. Please add any general comments you have on how useful you find summaries and why.

5. Please indicate how far along you feel you are with your dissertation as a percentage (this will be kept completely confidential, and used only to see if the usefulness of summaries is related to how specific your information needs are).

_____ %

“During” questionnaire

A copy of this questionnaire is to be completed after viewing three summaries for one article. It is intended to gauge your initial impressions of the summaries.

1. Would you read the article after viewing these summaries?

Yes No Don't know

Use these clouds as a reminder of what you have seen for the next questions:



Small cloud



Large cloud



Tag cloud

2. Which summary is most likely to encourage you to read the article? (please mark one)

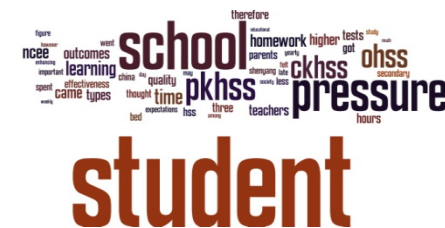
Small cloud Large cloud Tag cloud

3. Which summary did you like best? (please mark one)

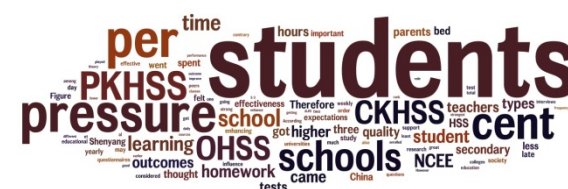
Small cloud Large cloud Tag cloud

“After” questionnaire – accompanying word cloud sheet

Small clouds



Large clouds



Tag clouds

academic academic-performance achievement activities
 class college course credit day distance entertainment family found groups hours identified
 important investigate learning marketing measure motivation music outcomes per
 performance personal profiles provide quality relationship reported research results significant
 spend spent student study study time table use university used
 variables watching week work

accomplish achievement attitudes average block college components computers
 effects factor goals grade long-range measured people points
 planning point practices priorities questionnaire relationship responses results
 sat score several single skills student subject tasks testing times
 time-management work

bei china ckhss colleges considered educational effectiveness enhancing expectations felt figure ping higher
 homework hours hss important improve school secondary shenyang society
 parents pkhss pressure quality school secondary shenyang society
 student study support teachers tests therefore thought types

“After” questionnaire

This questionnaire is to be completed after you have completed the task and viewed all summaries. It is intended to survey what you found useful about the summaries.

Please complete all the following questions in relation to the summaries you have just seen.

Part One – General questions about the summaries, considering all three articles at once. Letters refer to the attached sheet to serve as a reminder of summary types.

1. As a percentage, approximately how much of each summary did you read? (mark on line)

0-----→100%

a) Small cloud -----

b) Large cloud -----

c) Tag cloud -----

2. Did you feel that you could confidently accept or reject the article from the content of the summary?

a) Small cloud Yes No Don't know

b) Large cloud Yes No Don't know

c) Tag cloud Yes No Don't know

3. Assess your general impression of article content from each summary (mark on line).

Vague-----→ Detailed

a) Small cloud -----

b) Large cloud -----

c) Tag cloud -----

4. Assess how clearly you feel you understood the positioning and argument of the article from each summary (mark on line).

Vague-----→ Detailed

- a) Small cloud -----
b) Large cloud -----
c) Tag cloud -----

Part Two – More specific questions on how different summaries relate to selection of articles.

5. **Small clouds**

The sheet shows the small clouds seen for each article. Would you go ahead and read the article on the strength of this summary?

- a) Article One Yes No Undecided
b) Article Two Yes No Undecided

6. What did you find useful about this type of summary?

7. What did you find unhelpful about this summary?

8. Did you notice anything specific in these summaries that made you want to read the articles? (eg particular word/phrase)

9. **Large clouds**

The sheet shows the large clouds seen for each article. Would you go ahead and read the article on the strength of this summary?

- a) Article One Yes No Undecided
b) Article Two Yes No Undecided

10. What did you find useful about this type of summary?

11. What did you find unhelpful about this summary?
-
12. Did you notice anything specific in these summaries that made you want to read the articles? (eg particular word/phrase)
-
13. **Tag clouds**
- The sheet shows the tag clouds seen for each article. Would you go ahead and read the article on the strength of this summary?
- a) Article One Yes No Undecided
- b) Article Two Yes No Undecided
14. What did you find useful about this type of summary?
-
15. What did you find unhelpful about this summary?
-
16. Did you notice anything specific in these summaries that made you want to read the articles? (eg particular word/phrase)
-

Part Three – General questions on all summaries

17. Where you chose to read the article, which of these aspects of the summary contributed to your decision? (mark as many as appropriate)
- Prominence of key words Covered key topics New vocabulary
- Context Colour Background to topic
- Clear argument Other (please specify) _____
18. Please add any general comments you have on how useful you found these summaries and why.
-
-
-
-

Colour vs. black and white

