Testing an automatic organisation of retrieved images into a hierarchy

Mark Sanderson, Jian Tian, Paul Clough

Department of Information Studies, University of Sheffield,
Regent Court, 211 Portobello St,
Sheffield, S1 4DP, UK
(m.sanderson|p.d.clough)@shef.ac.uk

Abstract

Image retrieval is of growing interest to both search engines and academic researchers with increased focus on both content-based and caption-based approaches. Image search, however, is different from document retrieval: users often search a broader set of retrieved images than they would examine returned web pages in a search engine. In this paper, we focus on a concept hierarchy generation approach developed by Sanderson and Croft in 1999, which was used to organise retrieved images in a hierarchy automatically generated from image captions. Thirty participants were recruited for the study. Each of them conducted two different kinds of searching tasks within the system. Results indicated that the user retrieval performance in both interfaces of system is similar. However, the majority of users preferred to use the concept hierarchy to complete their searching tasks and they were satisfied with using the hierarchical menu to organize retrieved results, because the menu appeared to provide a useful summary to help users look through the image results.

1. Introduction

One process that users must perform when information seeking is to examine and interpret the search results. In most Information Retrieval (IR) systems, results are ranked in order of relevance to the query. However, if many search results are returned it can be difficult for the user to examine them all. In addition, reliably providing an intuitive summary of the search results is an obvious benefit to any user of an IR system. Hearst (1999) discusses various interface techniques for summarising results to make the document set more understandable to the user. These include: visualising the relationship of documents to the query, providing collection overviews and highlighting potential relationships between documents.

A variety of *clustering* techniques have been developed in IR to group documents. This can help users to browse through the search results, obtain an overview of their main topics/themes and help to limit the number of documents searched or browsed in order to find relevant documents (i.e. limit exploration to only those clusters likely to contain relevant documents). Two common variations are: (1) to group documents by associated terms (i.e. a set of words or phrases define a cluster and membership is based on its containing a sufficient fraction of a cluster's terms), and (2) to assign documents to pre-defined thematic categories (manually or automatically). Scatter/Gather (Cutting et al, 1992) and the Vivisimo[1] metasearch engine are an example of the former and Yahoo! Categories an example of the latter.

Organizing a set of documents automatically based upon a set of categories (or concepts) derived from the documents themselves is an obviously appealing goal for IR systems: it requires little or no manual intervention (e.g. deciding on thematic categories) and like unsupervised classification, depends on natural divisions in the data rather than pre-assigned categories (i.e. requiring no training data). In this paper we make use of such an approach for organizing search results called concept hierarchies (Sanderson & Croft, 1999; Sanderson & Lawrie, 2000). This simple method of automatically associating terms extracted from a document set has been successfully used to help users searching and browsing for documents (Joho, Sanderson, Beaulieu, 2004). In this simple method, words and noun phrases (called concepts) are extracted from passages of the top *n* documents and organized hierarchically based on document frequency and a statistical relation called subsumption.

Given the simplicity of this method and its success for document retrieval, in this paper we apply concept hierarchies to textual metadata associated with images for image retrieval and user test the resulting system. There are many instances of when images are associated with some kind of text semantically related to the image (i.e. metadata or captions). For example, collections such as historic or stock-photographic archives, medical databases, art/history collections, personal photographs (e.g. Flickr.com) and the Web (e.g. Yahoo! Images). Retrieval from these collections is typically supported by text-based searching which has shown to be an effective method of searching images (Markkula & Sormunen, 2000). To enhance such systems, various approaches have been explored to organize search results based on either textual and visual features (or a combination of both). A summary of related work is provided in section 2. In practice, given the proliferation of textual metadata, investigating methods to exploit this text (e.g. for organizing results) is beneficial.

The paper is ordered as follows: in section 3 we describe how we used concept hierarchies as a method for presenting image search results by displaying extracted concepts within a hierarchical structure. We describe the methodology and results of two user experiments to test the system and finally conclude.

2. Related Work

For image retrieval, clustering methods have been used to organize search results by grouping the top *n* ranked images into similar and dissimilar classes. Typically this is based on visual similarity and the cluster closest to the query or a representative image from each cluster can then be used to present the user with very different images enabling more effective user feedback. For example Park et al. (2005) take the top 120 images and cluster these using hierarchical agglomerative clustering methods

[1] http://vivisimo.com

Figure 1: Example fragment from generated menu for the query "church"

(HACM). Clusters are then ranked based on the distance of the cluster from the query. The effect is to group together visually similar images in the results.

Other approaches have combined both visual and textual information to cluster sets of images into multiple topics. For example, Cai et al. (2004) use visual, textual and link information to cluster Web image search results into different types of semantic clusters. Barnard and Forsyth (2001) organize image collections using a statistical model which integrates semantic information provided by associated text and visual features provided by image features. During a training phase, they train a generative hierarchical model to learn semantic relationships between low-level visual features and words. The resulting hierarchical model associates segments of an image (known as *blobs*) with words and clusters these into groups which can then be used to browse the image collection.

Approaches using only semantic information derived associated text have also been used to organize search results and to aid browsing. For example, Yee, et al. (2003) describe Flamenco, a text-based image retrieval system in which users are able to drill-down results along conceptual dimensions provided by hierarchically faceted metadata. Categories are automatically derived from Wordnet synsets based on texts associated with the images, but assignment of those categories to the images is then manual. Finally, Rodden et al. (2001) performed usability studies to determine whether organization by visual similarity is actually useful. Interestingly, their results suggest that images organized by category/subject labels or were more understandable to users that those grouped by visual features.

3. Building Concept Hierarchies

The approach of building a concept hierarchy proposed by Sanderson and Croft (1999) aims to automatically produce, from a set of documents, a concept hierarchy similar to manually created hierarchies such as the Yahoo! categories. The main difference being that concepts are in fact words and phrases (referred to as *terms*) found within the given set of documents and not categories defined manually. In their method of building concept hierarchies, word and noun phrases (called concepts) are extracted from retrieved documents and used to generate a hierarchy. Concepts are associated based on the set of documents indexed by the two concepts: the more documents two terms share, the more similar they are. However, concept hierarchies go beyond simple grouping of terms by discovering whether concepts are also related hierarchically. Document frequency and a statistical relation called subsumption is used to generate a hierarchy by detecting whether a parent term refers to a related, but more general concept than its children (i.e. whether the parent's concept subsumed the child's). Using document frequency (DF) to determine the semantic specificity of concepts is commonly used for weighting terms in IR based on Inverse Document Frequency (IDF).

With subsumption, concept C_i is said to subsume concept C_j when a set of documents in which C_j occurs is a subset of the documents in which C_i occurs. Or more formally, when the following conditions are held: $P(C_j|C_i) \geq 0.8$ and $P(C_i|C_j) < 1$. The assumption is that C_i is likely to be more general than C_j because, first, the former appears more frequently than the latter [13], and second, the former subsumes a large part of C_j's document set. Also they are likely to be related since they co-occur frequently within documents. The results can be visualised

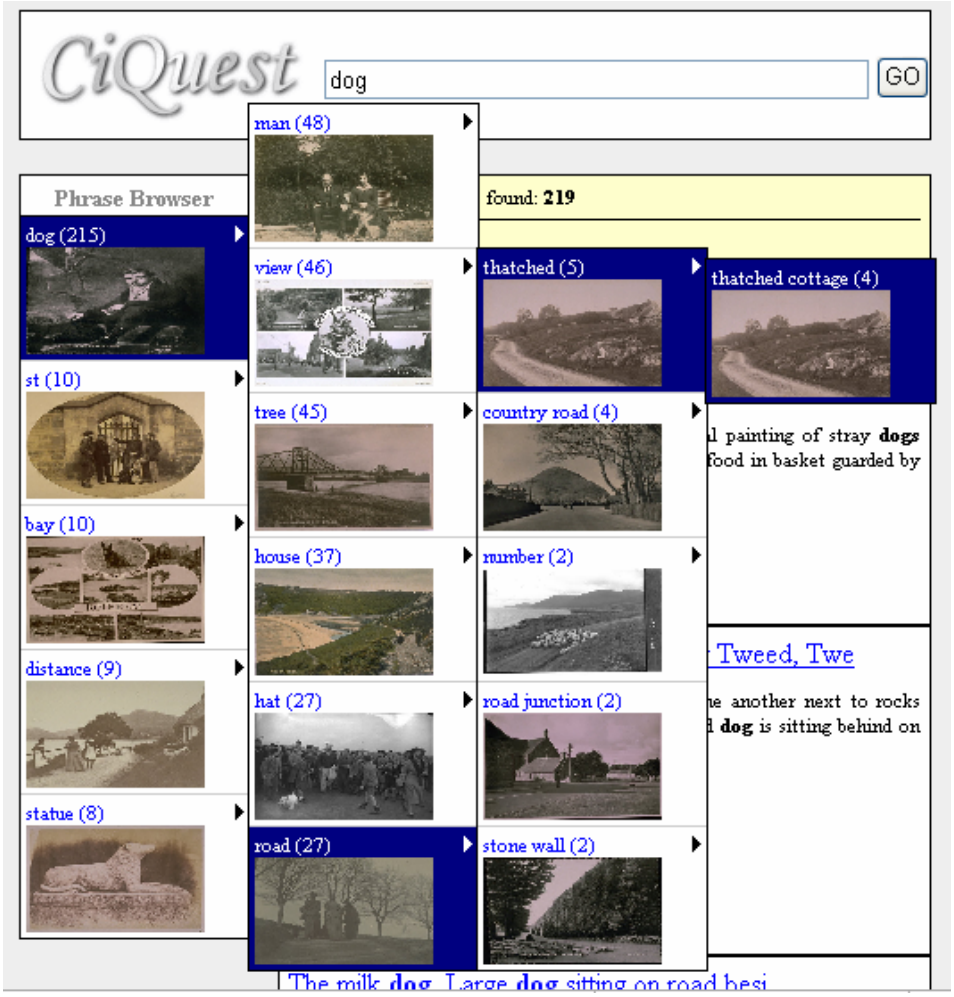

Figure 2: Example of the menu interface

using cascading menus where more general terms are placed at a higher level followed by related but more specific terms (Figure 1).

Sanderson and Croft analysed a random sample of parent-child relations and found that approximately 50% of the subsumption relationships within the concept hierarchies were of interest and that the parent was judged to be more general than the child. In particular, 49% of children were judged to reflect an aspect of the parent (a holonymic relation), e.g. actor is an aspect (or part) of a movie, 23% judged as a type of the parent (a hypernymic relation), e.g. a poodle is a type of dog, 8% judged to be the same as the parent, 1% as opposite to the parent, and 19% to be an unknown relation. We discuss relations commonly found using image captions in section 5. In summary, to generate a concept hierarchy for image browsing, the following steps are followed after an initial retrieval:

1. Extract concepts (words and noun phrases) from up to the top *n* image captions.
2. Compare each concept with every other concept and test for subsumption relationships.
3. Order concepts hierarchically based on DF scores (general to specific) and subsumption relation (concepts with no parent – no other concept subsumes - are top-level concepts).
4. Randomly select an image from the cluster to represent the cluster visually and create the menu.

For our image retrieval prototype, we used a version of the CiQuest system created to investigate user interaction with a standard textual document collection (Bernard & Forsyth, 2001). The system uses a probabilistic retrieval model based on the BM25 weighting function (Robertson et al 1995) to perform initial retrieval. A DHTML menu is generated dynamically representing the concept hierarchy, enabling users to interact with and browse the search results (Figure 1). The number in parenthesis is document frequency. A number of parameters can be adjusted in the prototype including:

1. *menu_depth*: maximum depth of menu;
2. *menu_height*: maximum height of menu;
3. *top_n*: number of documents to extract concepts from.

4. Experimental methodology

The current study is primarily concerned with evaluating the utility of the concept hierarchy menus to organise retrieved results and observe user interaction with the concept hierarchy menu based on a user-oriented task. To elaborate:

Figure 3: Example of the list interface

- evaluate the usability of concept hierarchy menus used in image retrieval from a user's perspective;
- obtain participants' perceptions of using concept hierarchy menus to group image retrieval results;
- gather participant's general impressions of menu interface (see Figure 2), compared with traditional list interface (see Figure 3); and
- analyse participants' searching behaviour with the concept hierarchy menu in image retrieval system.

4.1. Test Image Collection

The dataset used consisted 28,133 historic photographs from the library at St Andrews University[2]. All images are accompanied by a caption consisting of 8 distinct fields (short title, long title, description, location, date, photographer, notes and topic categories) which can be used individually or collectively to facilitate image retrieval. The 28,133 captions consist of 44,085 terms and 1,348,474 word occurrences; the maximum caption length is 316 words, but on average 48 words in length. All captions are written in British English and contain colloquial expressions and historical terms. Approximately 81% of captions contain text in all fields, the rest generally without the description field. In most cases the image description is a grammatical sentence of around 15 words. The majority of images (82%) are black and white, although colour images are also present. The dataset has been used for previous image retrieval experiments, the most notable being the ImageCLEF evaluation[3] campaign for cross-language image retrieval, see Clough, Mueller, and Sanderson (2005).

The methodology of the study was by means of conducting usability tests, including, task records, observation notes, pre- & post-session questionnaire and post-search interviews in order to get the perception of the participants. In the user test, each participant will be presented with two different version of the CiQuest interface and be asked to perform two user tasks on each.

4.2. Participants

A total of 30 participants were recruited for doing the user test. The majority of the participants (23) were graduate students of the Department of Information studies, University of Sheffield, and the rest were from other Departments of University. They consisted of 14 females and 16 males. The age of the participants ranges from 20 to 31 with an average of 25. All participated in the study as volunteers.

4.3. Experimental Tasks

Task one was designed as real life retrieval task, participants were required to search for images about a pre-specific topic using the CiQuest system with its different interfaces. In task two, participants were shown three photos taken from the St Andrews historic photographic collection and were required to find them using the CiQuest system with two different interfaces respectively. This task in real life can be described as,

[2] http://specialcollections.st-and.ac.uk/

[3] http://ir.shef.ac.uk/imageclef/

users trying to search for a specific image they have in mind; however, they do not know the exact keyword information to find it, so they need to describe the image by themselves. This task could be used to measure usability of experimental system, focusing on the effectiveness and efficiency.

In order to minimize order effects, users were shown either the menu interface first, or the list.

5. Results and Analysis

The results and analysis of current study are presented as follows.

5.1. Task One

In task one, each participant needed to work with both interfaces. Participants were asked to find 15 photos using CiQuest that were relevant to pre-designed topics. Based on their actual searching performance, participants were required to answer questions to evaluate the two different interfaces of the system. The participants were asked to work through 5 queries each. Results are presented in Table 1.

Mean score for task one	Menu	List
Av. number of pages user browsed	5	8
Av. number of queries type into system	1.6	3

Table 1: Mean score of five topics

As can be seen, in the list interface, users browsed more pages and entered more queries than when using the menu system. When participants use the list interface to search for photographs, they type the initial query into system and then at least examined one page of returned results to judge whether or not they need to reformulate their initial query. Based on author observation during the test, the majority participants were noted to browse at least two pages of results before they changed their query. So, if they change queries frequently, they must spend a lot times to view results. Therefore, in general the number of queries is proportional to the number of result pages.

When using the menu interface, the majority of participants spent time with the terms chosen for the menu as opposed to submitting a new query or going to view results page by page. The majority of participants used the menu interface usually to browse the first page of retrieved results in response to their initial query at first. Then if they could not find the relevant images they required, they prefer to view the concept menu before they went to the next page. They try to find appropriate terms on the menu to limit their initial retrieved results, and then they click term to browse associated results. If they could not find the photos, they went back to concept menu and tried other terms.

5.1.1. Questionnaire

Participants' general impressions of the two interfaces were gathered. Participants indicated on how easy or hard it was to find relevant images and how confident they were when locating images. The average time spent on completing this task was also shown in the table below.

As Table 2 shows, participants using the list interface spent more time on searching than using the menu interface a probable consequence of needing to enter more queries to complete their task. From observation of participants interaction with concept hierarchy menu, we can found the automatically generated concept hierarchy menu really helped users to narrow their result set down.

Task 1	Menu	List
Av. Time to complete task (min.)	10.2	12.4
How easy to judge relevance	4.0	3.2
How confident in judgements	4.1	3.8
Satisfied with the results	4.1	3.8

Table 2: Mean score of five topics

Also according to the table, the majority of participants thought it was easier to judge relevant images using the menu interface. The next question showed on the table was designed to evaluate how confident participants were with their relevant image choice. The mean score of using menu interface was 4.1, which slightly higher than mean score 3.8 of using list interface.

With information gained from the results of the experiments in Task one, we moved onto the second Task.

5.2. Task two

In task two, each participant again tested both the list and menu interfaces, with the aim of locating a "known item" image in the collection. All participants were asked to locate 3 images: half searched the menu interface first (referred to here as the menu group) and the other half used the list interface first (the list group). Results of the experiment are shown in Table 3

Task 2	Menu	List
Av. Time taken to find image	3.0	4.0
Av. number of result pages user browsed before finding the image	9.7	13.3
Av. number of queries	3.7	7.0
Success retrieval rate	91%	78%

Table 3: Mean score for task 2

As can be seen as with task one the average number of pages viewed and queries entered was smaller for the hierarchy interface than it was for the list, also (as before) the time users took to find the image on the menu system was shorter. What is more striking is the success rate of users in locating their known image: users were noticeably more successful in finding their target image with the menu system than they were for the list system. This result indicates that the concept hierarchy menu could provide some useful clues to help participants to find images. The concept hierarchy menu can improve retrieval effectiveness.

5.2.1. User behaviours

According to notes taken while observing users, the majority of participants in the menu group spent a lot of time browsing the menu. They seemed to prefer to view all parts of the menu, in order to find some similar images. They were particularly pleased when the required image was found with this strategy. Participants appeared to prefer searching through the menu than to re-formulate their query. It would appear that building a simple term hierarchy coupled with presenting that hierarchy in a quick browsing form is liked by users

6. Study findings

We analyzed the qualitative and quantitative results about they experimental system. By combining all results, some findings can be detected in this study.

The overall research aim of this study was to establish if the image retrieved results organized by automatically generated concept hierarchy menu is usable from the user perspective.

According to the task one result, image retrieval performance using menu interface was slightly better than using list interface. Although there was no significant difference between them, the results illustrated that the automatically generated hierarchy menu does support the image retrieval process. The concept hierarchy menu could group the image retrieved results by specific term related to the participants' initial query, in order to narrow the number of results returned to the screen. Based on the observation note, when participants used the menu interface, majority of them prefer to browse concept hierarchy menu choosing appropriate term instead of changing query or viewing a large number of results page by page. According to the evaluation questionnaire, the results illustrated that participants using menu interface were more satisfied with their task results than using list interface.

Secondly, from previous discussion of task two, although it was shown that there was no significant difference in retrieval performance between menu group and list group, using concept hierarchy menu can be seen as benefit to image retrieval process. The terms displayed on the concept hierarchy menu provided some useful clue for user to improve the successful rate on finding photos. Browsing concept hierarchy menu could be seen as providing an alternative choice for user to successfully find image, especially when participants' queries did not work.

Finally, based on the results of evaluation questionnaire, the majority of participants thought the menu interface is not as easy to use as list interface. However, the menu interface is easy to learn to use. All participants were never used the experimental system before. After the training session, they can easily learn to use it to complete two search tasks. Therefore, the learnability of the menu interface can be seen as acceptability. In addition, majority of participants gave the positive remark on concept hierarchy menu used in image retrieval. The satisfaction rate in menu interface was slightly higher than list interface. The majority participants were satisfied with using concept hierarchy menu to organize the retrieved results. They also mentioned that they prefer to use menu interface to retrieve image in the future.

However, some participants had a number of negative opinions in using menu interface. For example, two participants who favoured list interface mentioned that some terms displayed on the menu totally make them feel confused; they have no idea why these terms could be generated. Other participants also stated that some terms make them to the wrong path, result in waste a lot time and may sidetrack their original thought.

7. Conclusions

Overall the participants' impression of the experimental system CiQuest as image retrieval system was encouraging. They were satisfied with the search results and retrieval performance. Although both interfaces of experimental system had the similar capability to retrieve relevant images in response to users' query, majority participants prefer to use menu interface to organize their retrieved results in current study. Participants indicated that concept hierarchy menu could provide an intuitive preview for large numbers of retrieved results that gave them a better idea of the topics of image retrieved. So they can effectively narrow a lot returned retrieved results by choosing specific relevant topic, in order to avoid wasting so many time on browsing large numbers of results page by page. Participants also prefer to consider browsing concept hierarchy menu as an alternative way to help them successfully and effectively retrieve images, especially when their queries did not work well.

8. Acknowledgements

The work in this paper was part of a student Masters dissertation project conducted at the University of Sheffield and was also supported in part by the EU's BRICKS project IST-2002-2.3.1.12.

9. References

Bernard, K. and Forsyth, D. (2001) Learning the Semantics of Words and Pictures. In: *Proceedings of the Intentional Conference on Computer Vision*, vol 2, pp. 408-415.

Cai, D., He, Xiaofei., Li, Zhiwei., Ma, W-Y., and Wei, J-R. (2004) Hierarchical clustering of WWW image search results using visual, textual and link information. In: *Proceedings of the 12th annual ACM international conference on Multimedia*, pp. 952-959.

Clough, P., Mueller, H. and Sanderson, M. (2005), The CLEF 2004 Cross Language Image Retrieval Track, In: Peters, C., Clough, P., Gonzalo, J., Jones, G., Kluck, M. and Magnini, B. (Eds.) *Multilingual Information Access for Text, Speech and Images: Results of the Fifth CLEF Evaluation Campaign*, Lecture Notes in Computer Science, Springer, to appear.

Cutting, D.R., Karger, D.R., Pedersen, J.O., Tukey, J.W. (1992) Scatter/gather. A cluster-based approach to browsing large document collections. In *Proceedings of ACM SIGIR*

Hearst, M. (1999). User Interfaces and Visualization. In: Baeza-Yates, R. & Ribeiro-Neto, B. (eds.), *Modern Information Retrieval*, pp. 257-323. New York: ACM Press.

Joho, H., Sanderson, M., and Beaulieu, M. (2004) A Study of User Interaction with a Concept-based Interactive Query Expansion Support Tool. In: McDonald, S. & Tait, J. (eds), *Advances in Information Retrieval, 26th European Conference on Information Retrieval*, pp. 42-56.

Markkula, M. and Sormunen, E. (2000) End-use searching challenges indexing practices in the digital newspaper photo archive, *Information Retrieval*, 1, pp. 259-285.

Park, G., Baek, Y., and Lee, H-K. (2005) Re-ranking algorithm using post-retrieval clustering for content-based image retrieval, *Information Processing and Management*, 41(2), pp. 177-194.

Rodden, K., Basalaj, W., Sinclair, D., and Wood, K. (2001) Does Organisation by Similarity Assist Image

Browsing?, In: *Proceedings of the SIGCHI conference on Human factors in computing systems*, pp. 190-197.

Robertson, S.E., Walker, S., Beaulieu, M.M., Gatford, M. & Payne, A. (1995). Okapi at TREC-4. In: Harman, D.K. (ed.), *NIST Special Publication 500-236: The Fourth Text REtrieval Conference (TREC-4)*, Gaithersburg, MD. pp. 73-97.

Sanderson, M. and Croft, B. (1999) Deriving concept hierarchies from text In: *Proceedings of the 22nd ACM Conference of the Special Interest Group in Information Retrieval*, pp. 206-213.

Sanderson, M. and Lawrie, D. (2000) Building, Testing, and Applying Concept Hierarchies In: W. Bruce Croft, (ed.), *Advances in Information Retrieval: Recent Research from the CIIR*, Kluwer Academic Publishers, pp. 235-266.

Spärck Jones, K. (1972). A Statistical Interpretation of Term Specificity and its Application in Retrieval. *Journal of Documentation*, 28 (1), 11-21.

Yee, K-P., Swearingen, K., and Hearst, M. (2003) Faceted metadata for image search and browsing. In: *Proceedings of the SIGCHI conference on Human factors in computing systems*, pp. 401-408.