Exploring Large Digital Library Collections using a Map-based Visualisation

Mark Hall[1] and Paul Clough[2]

[1] m.mhall@sheffield.ac.uk
Department for Computer Science
University of Sheffield
Sheffield, UK
[2] p.d.clough@sheffield.ac.uk
Information School
University of Sheffield
Sheffield, UK

Abstract. In this paper we describe a novel approach for exploring large document collections using a map-based visualisation. We use hierarchically structured semantic concepts that are attached to the documents to create a visualisation of the semantic space that resembles a Google Map. The approach is novel in that we exploit the hierarchical structure to enable the approach to scale to large document collections and to create a map where the higher levels of spatial abstraction have semantic meaning. An informal evaluation is carried out to gather subjective feedback from users. Overall results are positive with users finding the visualisation enticing and easy to use.

1 Introduction

Access to current Digital Library (DL) services is primarily provided through the use of a search box and a query-response mode of interaction [20]. However, there are a number of situations when this is insufficient [21, 28] and where users would benefit from a richer user-interface to interact with [30]. For example, when users do not have clearly defined information needs [36]; when users have complex search tasks [33]; or when they want to gain an overview over a collection [15]. This has led to the design of browsing and exploration functionalities, including thesaurus-based search improvements [23, 31], document clustering [27] or the use of concepts arranged hierarchically in facets [12, 34]. All approaches have been shown to improve the user's search experience for exploration activities.

However, they are still primarily focused on improving the search experience and less on supporting overviewing and exploratory browsing tasks. Visualisations have been suggested as alternatives that would focus more on these two tasks and a number of visualisation techniques have been developed. The visualisations support the browsing experience [2], but struggle to scale to large collections and in the best cases support displaying a few thousand to tens of thousand of documents. This is either because there are computational limits

restricting the number of documents that can realistically be processed, or because there is a limit to the amount of information that can be displayed before the visualisation becomes unusable.

In this paper we present a novel approach to generating map-based visualisations for DLs that overcomes both these limitations. We achieve this by integrating hierarchical structure information, taken from a thesaurus that the documents have been mapped to, into the spatialisation process. This has three major advantages: first, it makes it possible to generate a map-based visualisation for collections with hundreds of thousands of documents; second, the hierarchical information means that the resulting map is not overloaded and remains usable with such a large collection; third: all structures shown on the map have an explicit semantic meaning, addressing issues of spatial-metaphor mis-use raised by [8]. The paper is structured as follows. Section 2 describes related work on collection overviews and information visualisation; Section 3 presents our approach to generating a 2D map-based visualisation of the concept or semantic space; Section 4 provides the results of an initial evaluation of the visualisation and Section 5 concludes the paper and provides directions for future research.

2 Background

2.1 Exploration of Digital Collections

In recent years approaches have been proposed for visualising document collections and providing overviews of the contents in order to support users with information exploration and discovery [14, 10]. The objective is to allow the user to gain an overview of the themes within an entire document collection which can provide a starting point for subsequent searches, especially if users are unfamiliar with the contents of a collection or are unable to formulate queries due to limited domain knowledge. A number of past studies have shown that visualising document collections can aid users with exploration [35, 2].

Overviews can vary from visualisations that display the individual documents in a collection and their relationships (i.e. the *document space*) to displays that show themes or topics associated with the contents of the documents (i.e. the *semantic space*). The design of overviews typically follows the principle of overview-then-details: providing zoomed out views of the collection and then allowing users to zoom in and inspect documents [32]. Various forms of visualisation have been used in the past, including spring-based visualisations [25], fish-eye views, 2D and 3D spatial visualisations [18, 24], network structures and hierarchical and tree-based visualisations of topics and documents [5, 3, 13, 14]. However, a common problem with visualisations is scaling them to deal with large collections and making them usable and understandable to users. For 2D or 3D spatialisation techniques [18, 24] the scaling limit is imposed by the computational complexity of the algorithms. For graph-based visualisations [25, 5, 3] the limit is defined by the amount of information that can be displayed before the visualisation becomes overloaded and unusable. The novel algorithm

presented in the next section overcomes both these issues through the creation of a hierarchically structured 2D map.

2.2 Spatial Metaphors for Information Visualisation

One type of visualisation that has been used in multiple domains is a 2D map which aims to spatially represent the contents of a document collection on maps or landscapes (a process known as *spatialisation*). These map-style visualisations are based on a spatial metaphor whereby similar documents (or themes) are clustered together and placed physically nearer to each other. They have been used for a range of applications, including providing thematic clustering of search results [1] and document collections [8], helping software developers make sense of large repositories of source code [17], allowing users to explore music libraries where "islands" on the map represent different genres or styles of music [26], and visualising user traffic on the Internet radio station last.fm [22].

Visualisations based on maps are often seen as a good metaphor in the sense that users of such systems are likely able to relate to the visualisation in a similar way they relate to maps derived from spatial data. However, Fabrikant et al. [8] argue that the spatial metaphor is often mis-used and that representing non-spatial information on maps or landscapes might not be as intuitive to users as one might think. For example the "islands" in [26], or contour lines in self-organizing maps [18], imply a spatial extent of the labelled concepts in the map, however all they actually represent is the spatial density of the individual documents. Thus documents belonging to a topic could lie outside of the "islands", breaking the spatial metaphor. To address this issue all spatial structures that our algorithm generates have a semantic meaning and observe basic geographic principles. Thus a boundary on our map denotes the border of a semantic area (in our case a topic in the thesaurus), just as a boundary on a geographic map denotes the border of a spatial object (such as a town or country). Similarly where areas in our map are contained within other areas, there is a part-of relationship between the contained object and the container, just as on a geographic map. This should ensure that people's common-sense assumptions about geography [7] are in line with how the semantic map works.

3 Hierarchical Spatialisation Algorithm

The novel Hierarchical Spatialisation Algorithm (HSA) presented in this paper requires as its input a thesaurus, a collection of documents, and a mapping of the documents into the thesaurus. From these it generates a hierarchical, semantic map using the six-step pipeline shown in Fig. 1. The first three steps (Tree pruning, Item pruning, Vectorisation) pre-process the thesaurus into the structure required for the main spatialisation algorithm. The next two steps (Spatialisation, Positioning) create the spatialisation of the thesaurus and the documents in the collection. The last step (Post-processing) creates the final map outline.

Fig. 1. High-level workflow of the hierarchical spatialisation algorithm

3.1 Pre-processing

The HSA requires that the thesaurus is a tree-structure and that documents are only found in the leaf nodes of the thesaurus. The pre-processing steps ensure that the thesaurus conforms to both criteria by first removing any multi-path structures and then pushing all documents into the leaf topics, creating new leaf topics where required. In addition to the structural requirements, the spatialisation algorithm also requires that all documents and topics have TFIDF vectors that define their position in the document space, and that are calculated in the third pre-processing step.

The restriction to tree-structures and that documents must be in the leaf nodes are limitations in the current implementation of the core spatialisation algorithm. However, Yu et al. [37] clearly demonstrate that thesauri with multiple paths to documents perform better in exploration tasks, as they increase the likelihood of the user finding any one path to the documents they are interested in. Thus work is currently ongoing to extend the core spatialisation algorithm to work with thesauri that are directed acyclical graphs.

Tree pruning ensures that the thesaurus is a pure tree, using a top-down, breadth-first algorithm that finds the highest location (closest to the root topic) for each topic in the thesaurus (Fig. 2). For each topic the pruning algorithm iterates over its child topics. If a child-topic has not yet been seen, then it is marked as "seen" (Fig. 2, b). If a child-topic is already marked as "seen" (Fig. 2, c) then the parent-child relationship is pruned, ensuring that each topic is placed at the highest location in the tree (Fig. 2, d).
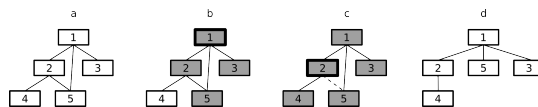

Fig. 2. Example of pruning a DAG hierarchy, the topic being processed is marked with a thick border. a) the original hierarchy where topic 5 has two parents; b) processing topic 1 and marking its children as "seen"; c) processing topic 2, topic 5 has already been "seen" and is thus removed from topic 2; d) the final thesaurus tree for spatialisation.

The highest location metric was chosen due to two criteria. First, a topic higher in the thesaurus is easier and faster to find. Second, a flatter thesaurus is better suited to the two-dimensional representation in the map. This heuristic was chosen as it worked best for our test collection and for other collections,

different heuristics might work better. With the future transition to supporting DAGs in the spatialisation, the question of which heuristic to use will become moot.

Item pruning After pruning the topic structure, the topic-document mappings also have to be pruned to ensure that each document belongs to only one topic. Items belonging to multiple topics, are removed from all non-leaf parent topics (Fig. 3, a). If after that step an document still has multiple parent topics, then it is assigned to the topic that is deepest in the hierarchy (Fig. 3, b). The justification being that a topic deeper in the hierarchy is likely to be more specific and thus provide a better fit for the document. If there are multiple parent topics at the same depth, then the document is randomly assigned to one. Only leaf topics may contain documents, thus for all non-leaf topics that contain documents a new child-topic is added with the original topic's label. All of the original topic's documents are then moved into the new child-topic (Fig. 3, c).
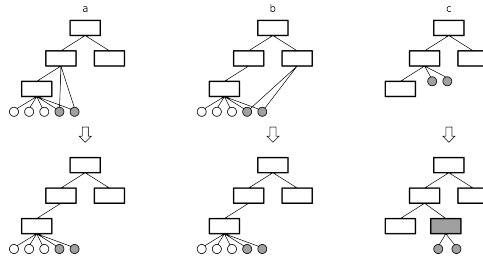

Fig. 3. Item pruning examples. a) pruning from non-leaf topics; b) pruning from higher-level topics; c) creating of an additional leaf node.

Vectorisation Items and topics are defined by TFIDF vectors calculated based on the documents' title, description, and subject keywords. To create the bag-of-words for each document, its title and description are tokenised and stop-worded, while the subject keywords are used as-is. Words that occur less than 5 times in the collection or in more than half of the documents in the collection are filtered. A single IDF vector is calculated based on all documents' bag-of-words. Then each document's TFIDF vector is calculated based on its bag-of-words and the IDF vector. For each topic the bag-of-words is created by combining the bag-of-words of all documents that belong to that topic or to one of the topic's descendants. The topic's TFIDF vector is then calculated based on the combined bag-of-words. As this leads to very large TFIDF vectors, for each topic only the 30 highest-scoring words are used for the final TFIDF vector. The threshold of 30 words was determined empirically, with a smaller number leading to indistinguishable topic vectors, while higher numbers created additional noise without improving topic distinguishability.

3.2 Hierarchical spatialisation

The core spatialisation algorithm uses a depth-first, bottom-up process to generate the documents' and topics' spatialisations (Fig. 4). Starting at the leaf nodes each topic is processed using the spatialisation algorithm described below. For leaf topics it is the documents that are spatialised, while for all other topics it is their child topics that are spatialised. After the documents or child-topics are spatialised the convex hull around them is calculated as the initial outline of the topic. This initial outline is then buffered in order to ensure that there is spatial separation between the topics. The buffered outline will be referred to as the topic's footprint. The centroid of the footprint is stored with the topic, as it is required for generating the final document placement.
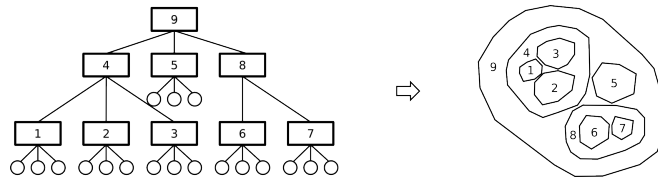
Fig. 4. Sample thesaurus hierarchy and the resulting spatialisation. The order in which the topics are processed is indicated by the numbers.

The algorithm achieves scalability to large collections through the ability to parallelise the process, as topics that are not in a parent-child relationship can be processed independently and concurrently. In the example in Fig. 4 the leaf topics 1, 2, 3, 5, 6, and 7 can all be processed in parallel. The same goes for topics 4 and 8. The only constraint in the parallelisation is that in order to be able to process a topic, all its child topics have to have been spatialised, as the spatialisation requires the child topics' footprints. Thus for example topic 4 can only be processed after topics 1, 2, and 3 have been spatialised. The scalability can be seen in the evaluation (Sect. 4), where a collection with approximately half a million documents is spatialised.

Spatialisation Items or topics are spatialised based on the TFIDF vectors calculated in the pre-processing step. By default the initial spatialisation is calculated using Multi Dimensional Scaling (MDS) [16]. However, if the number of elements to spatialise is larger than a given threshold α ($\alpha = 2000$ was used in the maps shown in this paper), then the elements are assigned co-ordinates randomly. While not ideal, it ensures that arbitrary numbers of elements can be processed in a realistic time-frame. At this point in the process both documents' and topics' locations are represented by points, the topics' spatial extents will integrated into the spatialisation at a later point.

One of the drawbacks of MDS is that if there are elements that are very different from the majority of elements to be spatialised, then this leads to a

degenerate spatialisation with a large cluster of points very closely positioned and the outliers at a distance (Fig. 5). To make the closely positioned points distinguishable on the map they have to be moved apart, however a simple linear scaling would result in maps with large amounts of empty space as the distance to the outliers would increase as well.


Fig. 5. Degenerate MDS with a large cluster of very densely located points and two outliers.

To address this problem the MDS spatialisation is transformed into a more compact representation (Fig. 6). First the Delaunay triangulation [6] for the MDS points plus a virtual point at $(0,0)$ is calculated using QHull[3] (Fig. 6, b). Due to the properties of the Delaunay triangulation, each point will be linked to its nearest neighbours in all directions. Then, for each point, the final co-ordinates p are calculated from the MDS co-ordinates m and the distance d from the $(0,0)$ point in the neighbourhood graph using equation 1. $\frac{m}{\|m\|}$ is the unit vector of m, a vector of length 1 that points in the same direction as the original vector m. Multiplying the unit vector by the distance d creates a concentric set of final co-ordinates for each point (Fig. 6, c), that maintains the qualitative spatial layout of the MDS spatialisation.

$$p = \frac{m}{\|m\|} \cdot d \tag{1}$$

Points are placed in the order of their distance d, thus the inner circle will be created first and then the outer circles added consecutively. The concentric spatialisation of the elements reduces the quantitative relationship between the point distance in the document vector space and the two-dimensional locations to a qualitative representation, however, the core principle that closer means more similar is maintained.
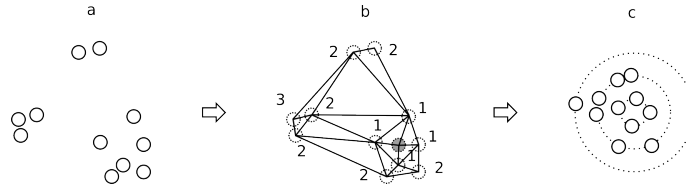

Fig. 6. The core spatialisation steps. a) the initial result of the MDS; b) the Delaunay triangulation with the neighbourhood distances; c) the final concentric spatialisation.

---

[3] http://www.qhull.org

When spatialising a set of topics, the topics' footprints have to be taken into account. To do this, each topic's footprint is placed so that its centroid is at point p. It is then checked if the footprint overlaps with other footprints that have already been placed and if it has then d is increased until it no longer overlaps. This results in a concentric spatialisation of the footprints where the footprints' boundaries touch, but do not overlap.

Final positioning To enable scalability the spatialisation algorithm processes each level in each sub-tree of the thesaurus independently. The result of the independent processing is that the topic footprints are not within the footprints of their parent topic and the documents are also not located within their topic's footprint (Fig. 7). To correct this the topics and documents are moved into their final position using the offset between their parent topic's footprint's current centroid and the centroid calculated when the footprint was created. This is performed in a top-down, depth-first manner, ensuring that each topic or document is moved only once, guaranteeing optimum performance.
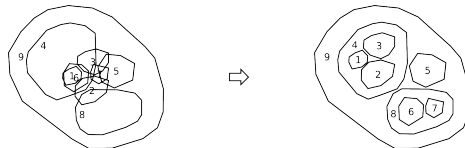
Fig. 7. On the left the placement of the footprints after the spatialisation, with footprints stacked above each other. At this point spatial containment does not indicate topical containment. On the right the spatialisation after final positioning and now spatial containment means topical containment.

3.3 Post-processing

After the spatialisation is completed, the topics' footprints will in most cases be touching their neighbours' footprints, which makes it harder to visually distinguish them. To correct this the algorithm removes the buffering added in the spatialisation process, ensuring sufficient distance between the topic outlines and creating the final map shown in Fig. 8.

4 Evaluation

Information visualisations are inherently complex to evaluate and multiple evaluation approaches exist [29, 9, 4]. We focus on user experience as opposed to evaluating aspects such as visual data analysis and reasoning, collaborative data analysis, or work practices [19]. The goal is to elicit subjective feedback and opinions on using the map in an undirected "exploring" context, to assess the
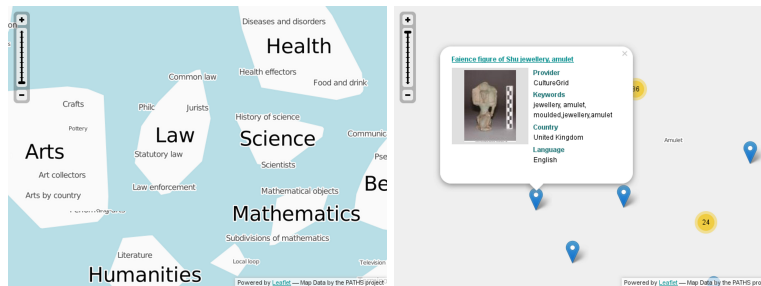
Fig. 8. On the left the initial overview over the map. On the right a zoomed-in showing individual documents and document clusters.

intuitiveness, utility, and usability of the map[19]. This cannot replace a full formal evaluation, but gives an initial indication of the visualisation's usability.

The HSA was evaluated using a collection of 547,780 documents taken from the English sub-set of Europeana[4], together with an automatically generated thesaurus [11] consisting of 27,049 topics covering all documents. Without parallelising any parts of the algorithm, the HSA required two days of processing on standard desktop PC, clearly demonstrating that it scales to large collections.

10 participants were shown the map and instructed to spend five to ten minutes exploring it. Leaflet[5] together with TileLite[6] were used to generate an interactive, digital map that provided the same interaction patterns as common on-line maps such as Google Maps. They were then asked to assess the system using a modified version of the the System Usability Scale (SUS) [4] that focuses on evaluating interactive digital maps, together with three quantitative questions derived from [19]. Additionally they were asked three qualitative questions.

Question	1st Quart.	Median	3rd Quart.
SUS Total Score	70	81.25	91.88
How understandable is the map?	4	4.5	5
Would you use the map again?	4	4.5	5
Did you generally know where in the map you were?	3	3.5	4

Table 1. The SUS score is out of a hundred, with 0 being "unusable" and 100 "perfect usability". The other questions are between "1" (bad) and "5" (good).

The participants spent an average 5 minutes 45 seconds exploring the map and a summary of their quantitative responses is shown in Table 1. Overall the results show that the participants felt comfortable using the map to explore the collection, although issues with orientation within the map remain. The

---

[4] The European Digital Library – http://www.europeana.eu

[5] http://leafletjs.com/

[6] https://bitbucket.org/springmeyer/tilelite/wiki/Home

qualitative answers support these conclusions, with comments including "The freedom of movement among topics" and "Overview of the topics", but also "Some text was overlapped" and "Need to double click to zoom in - took me a while to figure this out".

5   Conclusions

In this paper we describe a novel spatialisation algorithm that can scale to hundreds of thousands of documents and provides an intuitive map-based visualisation that can be used to hierarchically navigate themes in a document collection. This is possible by exploiting the hierarchical relationships between topics in a thesaurus that is attached to the documents. An informal evaluation shows that the resulting map succeeds in its goal of providing an enticing interface for getting an initial overview over the collection and exploring it. The 2D map would also be suitable for presenting collection overviews on mobile devices with limited screen space and when using a more tactile form of interaction. The visualisation has been deployed in the PATHS project[7], demonstrating its integration into a larger digital cultural heritage exploration system.

A number of areas to explore in future work remain. The foremost is to perform a more formal evaluation to determine the maps performance in the "overviewing" and "undirected exploring" tasks and comparing it to established interfaces such as faceted search or tag-clouds. A second major area is how to continuously update the map as the underlying collection changes, without having to re-calculate the whole map. Finally the cartographic styling of the map needs to be investigated to ensure that users understand it correctly.

6   Acknowledgements

References

1. Keith Andrews, Christian Gutl, Josef Moser, Vedran Sabol, and Wilfried Lackner. Search result visualisation with xfind. In *User Interfaces to Data Intensive Systems, 2001. UIDIS 2001. Proceedings. Second International Workshop on*, page 0050. Published by the IEEE Computer Society, 2001.
2. Marcus A. Butavicius and Michael D. Lee. An empirical evaluation of four data visualization techniques for displaying short news text similarities. *International Journal of Human-Computer Studies*, 65(11):931 – 944, 2007.
3. Matthew Carey, Daniel Heesch, and Stefan Rüger. Info navigator: A visualization tool for document searching and browsing. In *9th International Conference on Distributed Multimedia Systems (DMS)*, March 2003.

---

[7] http://www.paths-project.eu

4. A. Çöltekin, B. Heil, S. Garlandini, and S. I. Fabrikant. Evaluating the effectiveness of interactive map interface designs: A case study integrating usability metrics with eye-movement analysis. *Cartography and Geographic Information Science*, 36(1):5–17, 2009.

5. Chaomei Chen, Timothy Cribbin, Jasna Kuljis, and Robert Macredie. Footprints of information foragers: behaviour semantics of visual exploration. *International Journal of Human-Computer Studies*, 57(2):139 – 163, 2002.

6. B. Delaunay. Sur la sphère vide. *Izvestia Akademii Nauk SSSR, Otdelenie Matematicheskikh i Estestvennykh Nauk*, 7:793–800, 1934.

7. M.J. Egenhofer and D.M. Mark. Naive geography. In *COSIT'95: Conference on Spatial Information Theory*, volume 988 of *Lecture Notes in Computer Sciences*, pages 1–15. Springer-Verlag, 1995.

8. S. I. Fabrikant, D. R. Montello, , and D. M. Mark. The natural landscape metaphor in information visualization: The role of commonsense geomorphology. *Journal of the American Society for Information Science and Technology*, 61:253–270, 2010.

9. Camilla Forsell and Jimmy Johansson. An heuristic set for evaluation in information visualization. In *Proceedings of the International Conference on Advanced Visual Interfaces*, AVI '10, pages 199–206, New York, NY, USA, 2010. ACM.

10. Stephan Greene, Gary Marchionini, Catherine Plaisant, and Ben Shneiderman. Previews and overviews in digital libraries: designing surrogates to support visual information seeking. *Journal of the American Society for Information Science and Technology*, 51(4):380–393, March 2000.

11. Mark M. Hall, Clough Paul D., Samuel Fernando, Mark Stevenson, Aitor Soroa, and Eneko Aguirre. Automatic generation of hierarchies for exploring digital library collections. forthcoming.

12. Marti A. Hearst. Clustering versus faceted categories for information exploration. *Communications of the ACM*, 49(4):59–61, 2006.

13. Marti A. Hearst. *Search User Interfaces*. Cambridge University Press, New York, NY, USA, 1st edition, 2009.

14. Drahomira Herrmannova and Petr Knoth. Visual search for supporting content exploration in large document collections. *D-Lib Magazine*, 18(7/8), 2012.

15. Kasper Hornbæk and Morten Hertzum. The notion of overview in information visualization. *International Journal of Human-Computer Studies*, 69(7-8):509 – 525, 2011.

16. Joseph B Kruskal. Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis. *Psychometrika*, 29(1):1–27, 1964.

17. Adrian Kuhn, David Erni, and Oscar Nierstrasz. Embedding spatial software visualization in the ide: an exploratory study. In *Proceedings of the 5th international symposium on Software visualization*, SOFTVIS '10, pages 113–122, New York, NY, USA, 2010. ACM.

18. Krista Lagus, Samuel Kaski, and Teuvo Kohonen. Mining massive document collections by the websom method. *Information Sciences*, 163(1-3):135 – 156, 2004. Soft Computing Data Mining.

19. Heidi Lam, Enrico Bertini, Petra Isenberg, Catherine Plaisant, and Sheelagh Carpendale. Seven guiding scenarios for information visualization evaluation. Technical report, Department of Computer Science, University of Calgary, 2011.

20. Chern Li Liew. Online cultural heritage exhibitions: a survey of information retrieval features. *Program: electronic library and information systems*, 39(1):4–24, 2005.

21. Gary Marchionini. Exploratory search: From finding to understanding. *Communications of the ACM*, 49(4):41–46, 2006.

22. Daisuke Mashima, Stephen G. Kobourov, and Yifan Hu. Visualizing dynamic data with maps. In *Proceedings of the 2011 IEEE Pacific Visualization Symposium*, PACIFICVIS '11, pages 155–162, Washington, DC, USA, 2011. IEEE Computer Society.

23. David N Milne, Ian H Witten, and David M Nichols. A knowledge-based search engine powered by wikipedia. In *Proceedings of the sixteenth ACM conference on Conference on information and knowledge management*, pages 445–454. ACM, 2007.

24. Glen Newton, Alison Callahan, and Michel Dumontier. Semantic journal mapping for search visualization in a large scale article digital library. In *Second Workshop on Very Large Digital Libraries at ECDL 2009*, 2009.

25. Kai A. Olsen, Robert R. Korfhage, Kenneth M. Sochats, Michael B. Spring, and James G. Williams. Visualization of a document collection: The vibe system. *Information Processing & Management*, 29(1):69 – 81, 1993.

26. Elias Pampalk, Andreas Rauber, and Dieter Merkl. Content-based organization and visualization of music archives. In *Proceedings of the tenth ACM international conference on Multimedia*, MULTIMEDIA '02, pages 570–579, New York, NY, USA, 2002. ACM.

27. P. Pirolli, P. Schank, Marti A. Hearst, and C. Diehl. Scatter/gather browsing communicates the topic structure of a very large text collection. In *Proceedings of the SIGCHI conference on Human factors in computing systems: common ground*, pages 213–220. ACM, 1996.

28. Peter Pirolli. Powers of 10: Modeling complex information-seeking systems at multiple scales. *Computer*, 42(3):33–40, 2009.

29. Catherine Plaisant. The challenge of information visualization evaluation. In *Proceedings of the working conference on Advanced visual interfaces*, AVI '04, pages 109–116, New York, NY, USA, 2004. ACM.

30. Ramana Rao, Jan O. Pedersen, Marti A. Hearst, Jock D. Mackinlay, Stuart K. Card, Larry Masinter, Per-Kristian Halvorsen, and George C. Robertson. Rich interaction in the digital library. *Communications of the ACM*, 38(4):29–39, 1995.

31. A.A. Shiri, C. Revie, and G. Chowdhury. Thesaurus-enhanced search interfaces. *Journal of Information Science*, 28(2):111–122, 2002.

32. Ben Shneiderman, David Feldman, Anne Rose, and Xavier Ferré Grau. Visualizing digital library search results with categorical and hierarchical axes. In *Proceedings of the fifth ACM conference on Digital libraries*, pages 57–66. ACM, 2000.

33. G. Singer, U. Norbisrath, and D. Lewandowski. Ordinary search engine users carrying out complex search tasks. *Journal of Information Science*, 2012.

34. E. Stoica, Marti A. Hearst, and M. Richardson. Automating creation of hierarchical faceted metadata structures. In *Human Language Technologies: The Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL-HLT 2007)*, pages 244–251, 2007.

35. S. J. Westerman and T. Cribbin. Mapping semantic information in virtual space: dimensions, variance and individual differences. *International Journal of Human-Computer Studies*, 53(5):765 – 787, 2000.

36. Ryen W. White, Bill Kules, Steven M. Drucker, and M.C. Schraefel. Introduction. *Communications of the ACM*, 49(4):36–39, April 2006.

37. J. Yu, J.A. Thom, and A. Tam. Ontology evaluation using wikipedia categories for browsing. In *Proceedings of the sixteenth ACM conference on Conference on information and knowledge management*, pages 223–232. ACM, 2007.