Inf Retrieval (2014) 17:351–379
DOI 10.1007/s10791-014-9242-y

Evaluating hierarchical organisation structures for exploring digital libraries

Mark M. Hall · Samuel Fernando · Paul D. Clough · Aitor Soroa · Eneko Agirre · Mark Stevenson

Received: 31 January 2014 / Accepted: 2 July 2014 / Published online: 11 July 2014

Abstract Search boxes providing simple keyword-based search are insufficient when users have complex information needs or are unfamiliar with a collection, for example in large digital libraries. Browsing hierarchies can support these richer interactions, but many collections do not have a suitable hierarchy available. In this paper we present a number of approaches for automatically creating hierarchies and mapping items into them, including a novel technique which automatically adapts a Wikipedia-based taxonomy to the target collection. These approaches are applied to a large collection of cultural heritage items which is formed through the aggregation of other collections and for which no unified hierarchy is available. We investigate a number of novel user-evaluated metrics to quantify the hierarchies' quality and performance, showing that the proposed technique is preferred by users. From this we draw a number of conclusions as to what makes a hierarchy useful to the user.

M. M. Hall (✉)
Department of Computing, Edge Hill University, Ormskirk L39 4QP, UK
e-mail: Mark.Hall@edgehill.ac.uk

S. Fernando · M. Stevenson
Department of Computer Science, Sheffield University, Sheffield S1 4DP, UK
e-mail: s.fernando@sheffield.ac.uk

M. Stevenson
e-mail: mark.stevenson@sheffield.ac.uk

P. D. Clough
Information School, Sheffield University, Sheffield S1 4DP, UK
e-mail: p.d.clough@sheffield.ac.uk

A. Soroa · E. Agirre
IXA NLP Group, University of the Basque Country, 20018 Donostia, Basque Country, Spain
e-mail: a.soroa@ehu.es

E. Agirre
e-mail: e.agirre@ehu.es

Keywords Evaluation · Hierarchical structures · Exploratory search · Interactive information retrieval · Browsing

1 Introduction

There are many situations in which users of an Information Retrieval (IR) system may benefit from having documents organised into subject categories for browsing and exploration. For example, when users do not have clearly defined information needs (White et al. 2006), when attempting complex search tasks (Singer et al. 2012) or when they want to gain an overview over a collection (Hornbæk and Hertzum 2011). In such cases the provision of only a simple search box is insufficient (Marchionini 2006; Pirolli 2009). This is particularly relevant to digital libraries where rich user/information interaction is common and requires alternative methods to support users (Rao et al. 1995). The provision of browsing functionalities through thesaurus-based search enhancements (Milne et al. 2007; Shiri et al. 2002), document clustering (Pirolli et al. 1996) or the use of concepts arranged hierarchically in facets (Hearst 2006a; Stoica et al. 2007) have all been shown to improve the search experience.

To enable browsing, the items in a collection are typically mapped to a set of subject categories (e.g. a thesaurus or classification scheme), arranged for navigation, either hierarchically or as a set of facets. Traditionally this would have been done manually, creating a standardised, uniform subject categorisation for the collection. Due to the scale of modern, digital collections and the increasingly distributed nature of the collection creation process (Jörgensen 2004), creating a uniform subject categorisation requires significant manual and automatic post-processing (Yakel et al. 2007), which is not viable for many collections, particularly those at the big data scale.

Automatic creation of a hierarchy and the mapping between the items and the hierarchy offers a solution to this issue. This can involve the use of manually-created lexical resources such as WordNet (Navigli et al. 2003), automatically-generated hierarchies of concepts (or topics) derived from items in the collection (Sanderson and Croft 1999; Blei et al. 2003), or a combination of both (Stoica et al. 2007). However, there are a number of problems with such approaches. These include finding appropriate domain-specific lexical resources, the limited coverage of resources to items in a collection, the unfamiliarity of users with the concept labels, lack of cohesiveness between groups of concepts, and incorrect or unfamiliar parent–child relations between concepts. Recently, research has demonstrated how domain-specific thesauri can be mined from Wikipedia and used to improve IR, particularly for users who are unfamiliar with a domain (Milne et al. 2007). Despite this existing work, a number of open questions still remain. For example, how to successfully create hierarchies automatically, which existing lexical resources (if any) should be used and how should the quality of the hierarchies be evaluated.

This paper extends our previous work on comparing taxonomies for organising collections of documents (Fernando et al. 2012) and makes three major contributions: (1) a set of user-focused evaluation metrics that can be used to determine hierarchy and mapping quality; (2) a novel data-driven hierarchy creation algorithm that uses data derived from Wikipedia as an intermediary between the user and the data; and (3) the identification of attributes that suggest how hierarchies should be formed.

The paper is structured as follows. Sect. 2 describes related work; Sect. 3 describes resources and tools used in the experiments; Sect. 4 describes the hierarchies evaluated; Sect. 5 describes a on-line experiment to gather information about the quality of the hierarchies produced; Sect. 6 describes the novel Wikipedia-based hierarchy that we develop based on the results of the first experiment; Sect. 7 describes a task-based browsing activity to assess users preferences for a given hierarchy; Sect. 8 offers final discussion across all results and Sect. 9 concludes the paper and provides avenues for further work.

2 Related work

Hierarchies have been used to support the user in a number of IR tasks, including complex search tasks (Singer et al. 2012), collection browsing (Milne et al. 2007; Shiri et al. 2002), and hierarchical search facets (Hearst 2006a; Stoica et al. 2007). In this paper we focus on four key activities that hierarchies support:

1. Providing the user with an overview over the topics in a collection;
2. Providing the user with context information, when viewing a document or set of documents in the collection;
3. Support the user in unfocused exploration of the collection;
4. Organise the documents in the collection.

The goal in the evaluation metrics presented here are not to find the generically "best" hierarchy, but the one that is the most "useful" for the activities described above.

2.1 Generating hierarchies

Ideally, hierarchies for exploration are created manually (Rao et al. 1995; Rosenfeld and Morville 2002). While this is likely to lead to high-quality hierarchies, the process is too labour intensive to be feasible for large digital libraries. The manual process can also introduce language mismatches between the annotator and the user (Markkula and Sormunen 2000), making the resulting hierarchy harder to navigate and interpret correctly.

Automatically generating hierarchies is an alternative solution. Existing approaches can mostly be classified into those that assign items in the collection to the concepts in the existing hierarchy, and those that create the hierarchy first and then map the items to the hierarchy. Stoica et al. (2007) demonstrate an approach of the first kind, where they first map the documents in the collection to the WordNet hierarchy (Fellbaum 1998), and then create the sub-set of WordNet that covers the collection. A set of experts examined the resulting hierarchy and judged it to be useful. Navigli et al. (2003) use a similar approach which also was based on WordNet, but aimed at creating a full ontology with reasonable properties. Other researchers Milne et al. (2007) use Wikipedia to create a domain-specific hierarchy based on the links between individual articles, showing that a user-generated resource is useful, in contrast to expert-curated hierarchies like WordNet. Tang et al. (2006) demonstrate a generic algorithm that takes as input an existing item-to-hierarchy mapping, and adapts the hierarchy structure, using the similarity between the items to create a collection-specific veresion of the initial hierarchy.

While these approaches all require a pre-existing hierarchical resource to provide the relations between concepts, a number of alternative approaches are based only on the actual collection have also been proposed. Sanderson and Croft (1999) demonstrate such a data-driven approach, generating a hierarchy from the digital library's meta-data. Using term co-occurrence, they define a relaxed version of subsumption that states that, in most cases, two concepts are in a parent–child relationship. Using subsumption relation they then derive a complete hierarchy that covers the whole collection. In a similar approach Lawrie et al. (2001) use conditional probabilities to derive a hierarchy for a collection of documents. Another popular data-driven methodology, statistical topic modelling, has also been used to create hierarchies. For instance, Blei et al. (2003) describe a latent dirichlet allocation (LDA) variation that creates a hierarchical topic model. Alternatively, using a multi-branch clustering approach, Liu et al. (2012) automatically induce a hierarchy based on phrases extracted from the collection, which they combine with context knowledge derived from a knowledge base and web searches. Finally hierarchies based on sub-string matching have also been proposed (Anick and Tipirneni 1999; Nevill-Manning et al. 1999).

2.2 Evaluating hierarchies

One issue that arises with both manually and automatically created hierarchies is how to evaluate them (Lawrie et al. 2001). Historically they have primarily been evaluated using the following approaches:

- *Gold-standard* Typically used with automatically generated hierarchies, where the resulting hierarchy is compared to an existing (usually manually curated) hierarchy (Maedche and Staab 2002).
- *Criteria-based* The hierarchy is compared to a set of pre-defined criteria, such as consistency, completeness, or clarity (Gómez-Pérez 1996), which can be hard to evaluate automatically (Brewster et al. 2004).
- *Expert evaluation* The generated hierarchy is evaluated by a group of domain experts (Stoica et al. 2007).
- *Statistically* Lawrie et al. (2001) propose a number of statistical measures that can be used to automatically evaluate and compare hierarchies.

There are two issues with these evaluation approaches. First, they consider the hierarchy only from the view-point of the expert. This is sufficient if the target audience for the hierarchy is experts, but if the hierarchy is to be used by a wider user group, then non-expert based evaluation criteria are necessary.

The second issue is that these evaluation approaches consider the hierarchies essentially independently of the task that they were created to solve. Task-based evaluations (Pratt et al. 1999; Chen et al. 1999; Hearst 2006b; Yu et al. 2007; Wang et al. 2014) provide more information on how hierarchies and their navigation structures would perform in practice and overall there is a clear indication that systems that provide hierarchical navigation and query support outperform systems that do not. As these task-based evaluations tend to follow the "simulated work task" approach proposed by Borlund and Ingwersen (1997) their results also provide a strong indication as to how the hierarchies would perform in practice.

However, task-based evaluations are very time-, labour-, and resource-consuming (Toms et al. 2013), which has two effects. First, as Kelly and Sugimoto (2013) show, the number of participants in such experiments varies widely and a large number of different evaluation metrics are used, reducing the comparability of the results. Second, the complexity also means that in most cases only one or at most two hierarchies are evaluated. Where the goal of the evaluation is to determine whether and how a hierarchy can improve task performance, these are acceptable trade-offs. If, however, the goal is to choose which one of a set of potential hierarchies to use, then they represent a significant obstacle and a methodology that focuses on the comparability of large numbers of hierarchies is required.

In previous work (Fernando et al. 2012) we investigated whether simpler, user-evaluated metrics could be used instead of a full task-based evaluation, to enable the evaluation of larger numbers of hierarchies. The two metrics we investigated were whether the topics in the hierarchy were "cohesive" and whether the parent–child relationships were "sensible". By "cohesive" we mean that the items in a topic were closely related to each other and at the same time were clearly delineated from items in other topics. By "sensible parent–child" relationships we mean that, to the user, parent and child topics are obviously related, and that the type of relationship is also clear. Results showed the benefits of using Wikipedia as a basis for deriving topics and relationships even over the more expert-driven Library of Congress Subject Headings and WordNet Domains hierarchies, which in turn are better than the purely data-driven approach.

In this paper we extend this approach by evaluating two more metrics that directly address the core activities (overviewing, context, exploration, and organisation) listed above, and then comparing the results of the four metrics to a second, task-based evaluation to investigate whether the simpler metrics can be used as predictors for task performance.

3 Resources and tools

3.1 Europeana data-set

The base data-set on which the hierarchies are built is a collection of 547,780 cultural-heritage meta-data records taken from Europeana[1], acquired in spring 2011. Europeana is a web-portal to cultural heritage collections from over 2000 institutions based in Europe. Europeana records include information about a wide range of different types of media including paintings, films and archives.

The data-set is representative of digital cultural heritage collections in general, in that it has been aggregated from different sources, the amount of meta-data for each individual item is limited, and there is a large amount of variation in both the amount and semantic interpretation of meta-data from different sources. The data-set contains the English-language records provided to Europeana by 15 cultural heritage data holders in the United Kingdom. This may include records that mix in other languages as long as the record as a whole has been marked as English-language, however, a manual sampling analysis showed that these were very limited in number and unlikely to impact the data-processing. The number of records provided by the data-holders ranges between 4,144 and 125,562 records, with the largest six data-holders providing 86 % of all records. However, this has only a minimal impact on the variation, as the data provided by the individual data-provides is also highly variable in type and content.

We make use of three pieces of information from the Europeana metadata (see Table 1 for an example). The `dc:title`, `dc:description` and `dc:subject` fields that contain textual information describing each item. These fields were chosen since they are more informative than other fields in the meta-data and also tend to have been completed more consistently than other fields by the institutions that provide information to

[1] http://www.europeana.eu.

Table 1 Example Europeana item demonstrating the kind and amount of meta-data available to the hierarchy algorithms

dc:title	Clapham Common, Greater London
dc:description	A view showing Mount Pond.
dc:subject	Waterscape, public park, garden and park
thumbnail	

The thumbnail image was shown to the users in the evaluation, but is not used by any of the hierarchy algorithms

Europeana. 99 % of records have a `dc:title`, 74 % a `dc:description`, and 64 % at least one `dc:subject` value. The `dc:title` and `dc:description` provide short pieces of textual information about the item. The `dc:subject` field often links the item to an existing hierarchy, but not all providers have done this (i.e. the information is incomplete) and providers have also used a wide range of different hierarchies without documenting which one was used. For evaluation purposes, but not in the hierarchy creation process, we also used the thumbnail images provided by Europeana.

3.2 Wikipedia Miner

Wikipedia Miner (Milne and Witten 2008) is a freely available[2] Wikification tool which adds inline links to Wikipedia articles into free text. The software is trained on Wikipedia articles, and thus learns to disambiguate and detect links in the same way as Wikipedia editors. Milne and Witten (2008) report recall and precision of almost 75 % for the links generated by the tool.

4 Hierarchies

We investigate three Wikipedia-based approaches for generating hierarchies:

Wikipedia Taxonomy (WikiTax) focused on mapping items into the Wikipedia category hierarchy, Wikipedia Frequency (WikiFreq) as a data-driven approach using only the Wikipedia articles, and a DBPedia (DBPedia) based approach that maps items into the DBPedia ontology. As in Stoica et al. (2007) we use a LDA based approach to create a self-contained, data-driven hierarchy. We also compare these approaches to automatic mappings into the Library of Congress Subject Headings (LCSH) and WordNet Domains (WN Domains). These hierarchies are very different in their sizes, structure, topic generality, type of relationship between topics (*is-a* or other), and language. As there is no previous work on what kinds of hierarchies best support the activities identified above, the wide selection of hierarchies should ensure that we identify what type of hierarchy works for each of the activities.

4.1 Wikipedia taxonomy (WikiTax)

Wikipedia taxonomy (Ponzetto and Strube 2011) is a taxonomy derived from Wikipedia categories—a collaboratively-generated categorisation system that uses freely-chosen

[2] http://wikipedia-miner.cms.waikato.ac.nz/

- A.C. Cesena players
- A.C. Milan players
- A.C. Siena players
- A.S. Livorno Calcio players
- A.S. Roma players
- ACP magazine titles
- AEL Limassol players
- AIM clients
- ALCO locomotives
- AMC vehicles
- A Series of Unfortunate Events characters
- Abbadid
- Abdomen
    - Kidney
        - *Copper-alloy single-loop kidney-shaped buckle. Narrowed and off-set strap bar.*
- Abolitionism
- Abolitionist movements
- Abstract expressionism
- Abstraction
- Absurdist fiction
- Abugida writing systems
- Academic disciplines
    - Social sciences
        - Political science
            - *An anti-racist, sit-down protest where National Front literature is sold*
- ... 6848 further topics
- Zoroastrian history
- Zoroastrian texts
- Šumadija

Fig. 1 An extract from the WikiTax (Sect. 4.1) hierarchy. The extract clearly shows the very flat nature of the WikiTax hierarchy. All hierarchy extracts were generated automatically and show a selection of top-level topics, two sample branches, and for both branches the titles of two items (in *italics*). The automatic generation of the extracts ensures that they are accurate representations of the hierarchies overall structure (flat vs. deep)

keywords provided by contributors to Wikipedia. WikiTax is created by keeping the *is-a* relations between Wikipedia categories and discarding all others (Fig. 1). However, in some cases the relationships marked as *is-a* are not actual is-a relationships. As we have no way of automatically testing this, all relationships marked as *is-a* are retained. Into this hierarchy the Europeana items are then mapped by first applying Wikipedia Miner (see Sect. 3.2) over the Europeana items to find the relevant Wikipedia articles for each item. Each item is then linked to all Wikipedia categories that the item's articles belong to. While the approach relies on Wikipedia having articles that match the items' topics, it has been shown that Wikipedia in general has good coverage for most mainstream topics (Milne et al. 2007).

4.2 Wikipedia frequency (WikiFreq)

The Wikipedia Frequency approach was developed in previous work (Fernando et al. 2012). This approach also makes use of the links to Wikipedia articles in Europeana items that are identified by Wikipedia Miner (see Sect. 4.1). The articles linked to each item are used to form nodes in the hierarchy organised by how frequently they appear across all items in Europeana.

Let L be the set of articles that are linked to by Wikipedia Miner in at least one of the Europeana items. The frequency function $F : L \rightarrow \mathbb{N}$ gives the global frequency count for

- Alloy
- Chancel
    - Reredos
        - Arthur Blomfield
            - *St Werburgh's Church*
- Coin
- Copper
    - Brooch
        - *Brooch - Foot from a cast copper alloy Bow and Fantail brooch, dating to the...*
- ... 18 further topics
- Sherd
- Stoneware
- Watercolor painting

Fig. 2 An extract from the WikiFreq (Sect. 4.2) hierarchy. This hierarchy has a much more limited number of top-level topics and the topics labels generally fit the cultural heritage domain. The hierarchy extract was generated using the same algorithm as for the WikiTax hierarchy (Fig. 1)

Fig. 3 An extract from the DBPedia (Sect. 4.3) hierarchy. Similar in style to the WikiFreq hierarchy (Fig. 2), however the topic levels are less specific and the hierarchy generally flatter. The hierarchy extract was generated using the same algorithm as for the WikiTax hierarchy (Fig. 1)

- Activity
- AnatomicalStructure
    - Muscle
        - *Double-sided Pectoral Cross*
- Award
- Beverage
- ChemicalCompound
- Colour
- Currency
- Device
    - Weapon
        - *Ludworth, Moor Crescent, Bronze Age: Spot find - stone axe*
- ... 20 further topics
- Species
- Website
- Work

occurrences of the linked article in all items. Let $S \subset L$ be the set of articles in some item linked to by Wikipedia Miner. The articles in S are ordered by frequency, according to the function F with the most frequently occurring first, to produce an ordered list of articles $a_1, a_2, a_3 \cdots a_{|S|}$. This list of articles is then used to create a branch in the hierarchy $n_1 \rightarrow n_2 \rightarrow n_3 \cdots n_{|S|}$ such that each node, n_i, corresponds to a Wikipedia article, a_i, in the ordered set S. This branch is added to the hierarchy if it does not exist already. The articles' titles are used to label the nodes.

Finally, the hierarchy is pruned to remove nodes corresponding to articles that appear in fewer than 20 Europeana items, i.e. nodes n such that $F(n) < 20$. Additionally, where a node have more than 20 child nodes, only the 20 with the highest frequency are kept (Fig. 2).

4.3 DBPedia ontology (DBPedia)

The DBPedia ontology (Auer et al. 2007) is a small, shallow ontology that is manually created using information derived from Wikipedia. Unlike the other hierarchies, DBPedia is a formalised ontology, including inference capabilities. The ontology provides the instances of each ontology class, i.e. the set of Wikipedia articles pertaining to this class (Fig. 3). To map Europeana items to DBPedia classes, we first apply Wikipedia Miner to

Fig. 4 An extract from the LDA (Sect. 4.4) hierarchy. Clearly shows that the hierarchy is much deeper and also illustrates the labelling issues with the same label used at different levels. The hierarchy extract was generated using the same algorithm as for the WikiTax hierarchy (Fig. 1)

- Britain
- Brooch
  - Design
    - Design
      - Kitchen
        - *Victoria Buildings, Bath Street, 1976*
- Cash
- Coin
  - Coin
    - Elizabeth I
      - Roman
        - Measure
          - *Jetton*
- ... 3 further topics
- Number
- Vessel
- View

find the relevant Wikipedia articles for the item (as in Sect. 4.1), and then link the item to the classes these articles belong to.

4.4 Latent dirichlet allocation (LDA)

Latent dirichlet allocation is a soft-clustering algorithm which discovers so-called latent topics in a document collection. LDA also discovers the topics which are relevant for each document. LDA is an example of a data-driven, fully self-contained hierarchy. Previous approaches have used flat clusters (Stoica et al. 2007), but we use a recursive divide-and-conquer approach on top of LDA to create a full hierarchy (Fig. 4). We created the LDA-based hierarchy as follows:

1. LDA needs to decide before-hand the number of topics to be discorevered. We set this number ($topic\_n$) automatically depending on the number of items in the collection ($|C|$):

$$topic\_n = \min\left(9, \frac{|C|}{30}\right) \quad (1)$$

2. Pre-process each item to prepare the bag-of-words representation required by LDA. The text was first lematised and tagged with part-of-speech information using Freeling (Padró et al. 2010), a multilingual NLP tool. Nouns and adjectives were selected as the bag-of-words representation. For 1,884 items this approach produces empty bags-of-words, which means that those items cannot be processed into the hierarchy.
3. Run LDA to discover $topic\_n$ latent topics and assign all items to those topics. LDA returns an topic distribution for each item, so we assign each item to its highest-ranked topic.
4. LDA also provides a topic-word distribution, which is used to select the highest-ranked word as the topic's label.
5. Split the collection by topic into a set of sub-collections and recursively apply the algorithm to each sub-collection. Note that the number of topics ($topic\_n$) is updated in each iteration.
6. Stop if a sub-collection has fewer than 60 items or if the depth of a branch is larger than 10.

- Accidents
- Accidents
  - Fires
    - *Firemen at work at a midnight fire Title Series: London life by night (10513-60)...*
- Administration
- Administration
- Adolescence
- Aerial views
- Aesthetics
  - Romanticism
    - *Hull of a boat in dry dock. Based on metaphorical subject of loneliness in urban...*
- ... 508 further topics
- Wood carving
- Workshops
- Writing

Fig. 5 An extract from the LCSH (Sect. 4.5) hierarchy. Clearly shows the very flat structure of the hierarchy. The hierarchy extract was generated using the same algorithm as for the WikiTax hierarchy (Fig. 1)

The parameters used in the topic number calculation (9 and 30 in Eq. 1) and the stop condition (60 items min. and max. depth of 10) were determined empirically, by visualising the resulting hierarchy. Topics of around 30 items provided the best experience, with up to 60 items being acceptable from a user perspective. Similarly a minimum of 9 topics at any level provided the best balance between clearly distinct topics and an overly deep hierarchy. Finally the maximum branch depth of 10 was added to ensure that the algorithm always terminates.

We used LDA as the topic modelling algorithm for two reasons: (1) it is a state-of-the-art approach that scales well to the collection size and (2) it has been shown to create cohesive topics (Fernando et al. 2012; Chang et al. 2009). Also LDA has been used to successfully improve result quality in IR (Azzopardi et al. 2004; Wei and Croft 2006) tasks. Although only the highest-ranking item–topic assignment is used, discarding some useful information, the speed and cohesion of the resulting topics nevertheless make LDA a suitable approach.

4.5 Library of Congress (LCSH)

The Library of Congress Subject Headings comprises a controlled vocabulary maintained by the U.S. Library of Congress for use in bibliographic records. LCSH is widely used by libraries to organise their collections as well as for organising materials online. The Europeana item's text is lemmatised using Freeling (Padró et al. 2010). The text is compared to the category labels for the LCSH concepts. If the text contains any of the category labels then the item is matched to these categories. If more than one matching label is found, then the longest matching label is used for the mapping (Fig. 5).

4.6 WordNet Domains (WN Domains)

WordNet Domains (Magnini and Cavaglia 2000) comprise a set of 164 domain labels which have been semi-automatically assigned to each of the synsets in WordNet. The domain labels group together words from different syntactic categories (e.g. nouns and verbs), and also may group together different senses of the same word and thus reduce

- Applied science
- Factotum
    - Person
        - *Portrait of a Man*
- Free time
- Humanities
    - Paranormal
        - Occultism
            - *penny; denomination toy coins; Series Magician's Money; subseries DEMON...*
- Pure science
- Social

Fig. 6 An extract from the WN Domains (Sect. 4.6) hierarchy. The extract demonstrates the very small nature of the hierarchy, making it possible to show all top-level topics. The hierarchy extract was generated using the same algorithm as for the WikiTax hierarchy (Fig. 1)

polysemy. Note that Stoica et al. (2007) used the WordNet taxonomy, which has been questioned for being unintuitive for regular users (Horvat et al. 2012). We decided to use WN domains instead, which is simpler, smaller and, to our believe more intuitive for lay users.

Yago2 is used as an intermediate vocabulary for the mapping process. Yago2 (Hoffart et al. 2011) is a knowledge base derived from Wikipedia with more than 10 million entities, and each entity in Yago2 is linked to a WordNet 3.0 synset. We also used a mapping from WordNet 3.0 synsets to WordNet Domain labels as provided by the Multilingual Central Repository (MCR) (Atserias et al. 2004). To perform the mapping, the first step is linking Europeana items to Yago2 entities using the Freeling for lemmatisation and longest possible match approach as used for the LCSH mapping in Sect. 4.5. The Europeana items are then mapped to the WordNet Domain labels via the Yago2 entity-to-synset and the MCR synset-to-WordNet Domains mappings (Fig. 6).

4.7 Hierarchy filtering

The complete set of 547,780 Europeana items were processed by each of the hierarchy algorithms. Due to different approaches used by each algorithm, the resulting hierarchies cover different, but overlapping, parts of the data-set. To ensure that the hierarchies and mappings are comparable and the evaluation results valid, the experiments were run on the sub-set of the collection that is covered by all six hierarchies, a collection of 8,179 items. Topics in each hierarchy that did not contain any items from that set of shared items were pruned.

While the pruning step reduces the size of the collection that the experiments are run on, it is necessary to ensure the comparability of the different hierarchies' results. If no filtering were applied and the experiments showed significant differences between the hierarchies, it would always be uncertain whether the differences were due the differences in the hierarchy algorithms or due to differences in the hierarchies coverage or the processed items' meta-data. One hierarchy algorithm might process items with very poor meta-data that are ignored by the other algorithms, leading to a poor mapping into the hierarchy, and thus poor evaluation results. While the amount of coverage a hierarchy achieves is an important factor for choosing a hierarchy, by applying the filtering step we can separate the hierarchy quality and coverage aspects. The experiment results thus show how the hierarchies perform on the same data and these results can then be combined with the coverage to determine the most applicable hierarchy for a given use context.

Table 2 Hierarchy statistics before and after filtering to the shared item-set

Hierarchy	Pre-filtering		Post-filtering					
	#Topics	#Items	#Topics	#Root top.	Depth		Children	
DBPedia	273	178,312 (32.6 %)	105	20	2	4	2	11
LCSH	285,238	99,259 (18.2 %)	1,043	174	4	18	1	20
LDA	22,494	545,896 (99.6 %)	1,828	9	5	10	1	9
WikiFreq	502	66,558 (12.2 %)	211	19	2	6	1	21
WikiTax	121,359	275,359 (50.4 %)	4,036	1,798	2	12	1	44
WN Dom.	170	308,687 (56.5 %)	143	6	3	4	3	21

Depth and children are reported as median / maximum

Table 2 lists descriptive statistics for the unfiltered and filtered versions of each hierarchy. LDA covers almost the whole collection. WN Domains has the next highest coverage, while WikiFreq has the lowest coverage (12 %). The hierarchies cover a number of different styles: WikiTax is wide and quite shallow, LCSH is also wide, but deeper, LDA is narrow and deep, DBPedia and WN Domains are quite small. The wide variety of hierarchy shapes is intentional, as we were interested in investigating how these impact on the hierarchies perceived “usefulness”. The fact that for a number of hierarchies the median number of child topics is 1 is an artefact caused by the filtering process which prunes those parts of the hierarchies that do not cover the shared set of 8,179 items.

5 Hierarchy comparison

As stated above, the first of the two experiments presented in this paper was designed as an extension of previous work (Fernando et al. 2012), where we evaluated the hierarchies to determine how cohesive concepts were in the hierarchy and whether the methods produced logical relationships between concepts (as judged manually). This experiment extends the previous work to analyse the hierarchies for two additional aspects: whether they are perceived to provide an overview of the collection (Sect. 5.1) and whether individual items are “well-placed” in the hierarchies (Sect. 5.2), evaluating the key activities #1 (“provide an overview”) and #2 (“provide context”) respectively.

Testing hierarchies using these very specific aspects was chosen as an approach, because we wanted to evaluate a large number of hierarchies using non-expert users. This ruled out using a larger task-based evaluation setup, as the resource-commitments, primarily time and participants, required to get valid results severly limit the number of hierarchies that can be evaluated in parallel. As there was no existing work on what non-expert users preferred in a hierarchy, we wanted to investigate evaluation methods that did not require severly limiting the number of hierarchies tested. Additionally in a task-based experiment the evaluators’ responses will be influenced not only by the hierarchy, but also by their interest in the task and the collection. The second experiment (Sect. 7), where we use a task-based setup, shows that this approach can result in evaluation results that are harder to draw any significant conclusions from.

The goal of evaluating hierarchies for the use with novice, non-expert users also meant that gold-standard-, criteria-, and expert-based evaluation setups were not applicable, as these are all based on expert evaluation. Since there was no literature to show that expert evaluations provided information on how useful a hierarchy could be for novice, non-expert users, a setup that allowed us to test with non-expert users was required. Splitting the evaluation into the individual aspects makes it possible to create an experiment that is sufficiently in-depth for the aspects tested and at the same time is sufficiently short to attract a significant number of evaluators.

The experiment was designed as an on-line experiment using our own experiment support software (Hall and Toms 2013). It consisted of three parts: an initial set of background questions (age, gender, first language); the first research question investigating the hierarchies' overviewing capabilities; and the second question investigating the item placement. Participants were recruited from staff and students at Sheffield University via a central mailing list with approximately 20,000 subscribers. No incentives were offered for participation. A total of 881 people started the experiment and 288 completed it (32.6 % completion rate). After filtering participants who did not complete the experiment or were not first language English speakers a total of 225 participants remained. Filtering was conducted post-experiment to allow for a future comparison of sub-groups within the data-set. Of the 225 participants 136 were female and 89 male. Slightly fewer than half (106) were between 18 and 25, 103 are split relatively evenly between 26 and 55, and the remaining 16 are over 55.

5.1 Overviewing task

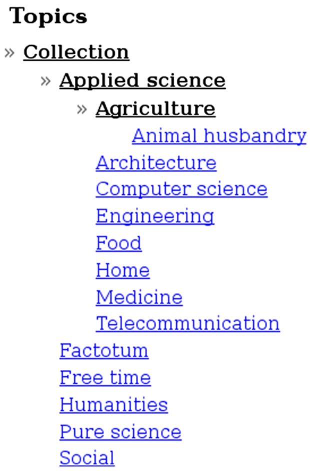
Participants were asked to complete the overviewing task before the item placement task. Organising the tasks in this way removes the risk of experience gained during the second task influencing performance in the first one.

5.1.1 Setup

The task itself was presented through two pages. On the first page the participants were shown one of the six hierarchies using our hierarchy browser (Fig. 7). Participants were automatically allocated one of the hierarchies in order to ensure a balanced distribution of participants to hierarchies. They were instructed that they had come across an unknown collection and should spend two minutes exploring the hierarchy in order to develop an overview of what is in the collection. Participants were given an explicit time-limit of two minutes after which the experiment automatically moved on to show the second page containing the following questions:

1. Q1: *How much of the collection do you believe you explored?* (0–100 % in 10 % steps)
2. Q2: *Please rate how good an overview over the collection you got* (7-point semantic differential; *good–bad*)
3. Q3: *Please rate how organised you felt the collection was* (7-point semantic differential; *organised–random*)
4. Q4: *Please rate how understandable the collection was* (7-point semantic differential; *understandable–not understandable*)
5. Q5: *Please rate how familiar you are with the topics covered in the collection* (7-point semantic differential; *familiar–unfamiliar*)
6. Q6: *Please rate how confident you are about what you would expect to see in the various parts of the collection* (7-point semantic differential; *confident–not confident*)

The percentage explored is coded $[0, 1]$ in 0.1 steps. The semantic differentials are coded from -3 (negative statement) to +3 (positive statement).



Topics

» Collection
 » Applied science
  » Agriculture
   Animal husbandry
  Architecture
  Computer science
  Engineering
  Food
  Home
  Medicine
  Telecommunication
 Factotum
 Free time
 Humanities
 Pure science
 Social

Fig. 7 Hierarchy browser used in the first experiment for the "Overviewing Task". Clicking on a topic would show / hide its child topics. No items were shown to avoid any influence from the item placement on the overviewing evaluation

In addition to these questions the hierarchy browsing interface also logged every interaction between the participant and the interface and the total amount of time spent browsing the hierarchy. From the interaction data three metrics were derived for each hierarchy: the total number of clicks (Clicks); the fraction of topics either directly selected (% Viewed) or where an ancestor was selected (% Ancestor Viewed).

Participants were only shown the hierarchy and not the items themselves. This was done for two reasons. First, the goal of the whole experiment was to test whether individual hierarchy characteristics could be tested separately, which is not possible when both the hierarchy and the items are shown, as that mixes the hierarchy quality with the item-placement quality questions. Second, the second task in this experiment was aimed at evaluating the item-placement and showing a participant the item-placement for one hierarchy and then asking them to evaluate the same placement for a number of hierarchies could potentially introduce a judgement bias.

The time limit of two minutes was determined using a small pilot study where we found that on the smallest of the hierarchies (DBPedia) in two minutes it was possible to explore all concepts. Thus a longer time period would favour DBPedia. In addition, in a pilot study participants stated they lost interest in the task after more than two minutes. While the time limit means that participants are unable to completely explore the hierarchy they are shown, it ensures that the results reflect a more realistic scenario. In most cases users do not explore everything before making a decision, they explore enough to satisfy themselves that they have a good-enough understanding and then decide whether the collection is of interest to them or not. Thus while the participants do not have a perfect understanding of the hierarchies in order to provide perfectly grounded responses, their level of understanding and thus their level of responses is more realistic.

5.1.2 Results

Table 3 summarises the results of the overviewing task. A Kruskal–Wallis test shows no statistically significant influence of hierarchy on the time spent ($\chi^2 = 7.37$, $df = 5$, $p = 0.19$), thus any differences in the results can be ascribed to the hierarchies and not the time participants allocated to the task. The number of clicks shows statistically significant

Table 3 Overviewing task results

Metric	DBPedia			LCSH			LDA		
Q1. Explored	.3	.5	.6	.1	.1	.1	.2	.3	.4
Q2. Overview	−1	1	2	−2	0	1	−2	− 1	1
Q3. Organised	0	1	2	−1	1	2	−2	−1	0
Q4. Understandable	−1	1	2	0	1	2	−3	−2	−1
Q5. Familiar	**1**	**2**	**2**	0	1	2	−2	−1	1
Q6. Confident	**0**	**1**	**2**	−1	1	2	−3	−2	1
Clicks	16.5	29.0	51.5	12.0	24.5	37.5	13.5	30.0	59.8
% Viewed	.081	.152	.257	.005	.012	.019	.005	.011	.021
% Ancestor viewed	.176	.352	.562	.015	.032	.050	.051	.175	.402

Metric	WikiFreq			WikiTax			WN Domains		
Q1. Explored	.2	.3	.4	0	.1	.1	.3	.5	.7
Q2. Overview	0	1	2	−3	−1	−0.5	**1**	**2**	**2**
Q3. Organised	−2	1	2	−1	2	3	**2**	**2**	**2.5**
Q4. Understandable	−1	1	2	−2	1	2	**1**	**2**	**3**
Q5. Familiar	−1	1	1.5	−1	−0.5	1	1	1	2
Q6. Confident	−1	0	1.5	−2	−0.5	1	**0**	**1**	**2**
Clicks	12.5	24.0	37.5	11.0	18.0	26.0	26.0	27.0	46.0
% Viewed	.033	.061	.097	.001	.002	.003	.07	.104	.167
% Ancestor viewed	.14	.284	.415	.003	.051	.053	.42	.741	1

All results are 1st quartile/median/3rd quartile. Second row (Explored) contains numbers from 0 to 1. The rest of rows from 2 to 6 contain a number between −3 (negative statement) and 3 (positive statement), with best numbers in bold. Clicks report absolute numbers. The last two rows report percentages from 0 to 1

differences ($\chi^2 = 35.46$, $df = 5$, $p < 0.001$). Pairwise Wilcoxon rank-sum tests show that the click counts (Clicks row in the table) come from three different groups: the first gets the most clicks and includes DBPedia, LDA, and WN Domains, which are all three clearly hierarchically organised; the second groups LCSH and WikiFreq; WikiTax with its extremely high number of top-level topics forms the third group.

This split is unexpected, as a simple binary division into deeper (DBPedia, LDA, WN Domains, WikiFreq) and flatter (LCSH, WikiTax) hierarchies has been expected. The flatter hierarchies were expected to have less clicks, as more of the hierarchy could be explored by scrolling. A possible cause for this triple-split is that the combination of different topic types at the same level in the LCSH and WikiFreq hierarchies has an impact. Both LCSH and WikiFreq mix conceptual and instance topics at the same level (e.g. "Symbolism" and "Table" next to each other in LCSH, or "Flint" and "Greater London" in WikiFreq) and this unclean structure impacts how the participants explored the collections.

On the core questions of how good an overview a hierarchy provides, how well organised it is, and how understandable it is (questions 2, 3 and 4 in the table), WN Domains outperforms the other hierarchies. DBPedia also performs quite well and interestingly in the "familiar" question is slightly better than WN Domains. The difference is not statistically significant, but might indicate that the language used in DBPedia is more familiar to the participants than the WN Domains language, although it might also simply be due to a sampling bias.

As had been expected the broad top-level hierarchies (LCSH, WikiTax) did not give as good an overview as the narrower hierarchies (DBPedia, WikiFreq, WN Domains). The WikiTax results highlight the problem of having too many top-level topics, as the results show that it does not give a good overview, even though it is rated as being well organised.

The LDA hierarchy struggles in all aspects, potentially because it is too narrow, but more likely because the topic labels are too simple and do not give a good overview over what can be found in the hierarchy. These results are in line with Stoica et al. (2007) who also show that the manually created hierarchies are seen as clearer.

Comparing the participant's self-evaluation of how much they explored (question 1) with the two fraction of topics metrics (two bottom rows in the table) clearly indicates that participants extrapolate from the topic to its children. Their self-evaluation is much closer to the "% Ancestor" metric than to the absolute number of topics selected. It seems that the participants are making assumptions about what lies below the topics they have explicitly seen and are making their judgements of how much they have explored on this basis. This re-enforces the hierarchy design rule that a good hierarchy must start with general topics and become more specific as the user drills down. It is also in line with Sanderson and Croft (1999) who state that where this relationship structure breaks, the users struggle with using the hierarchy.

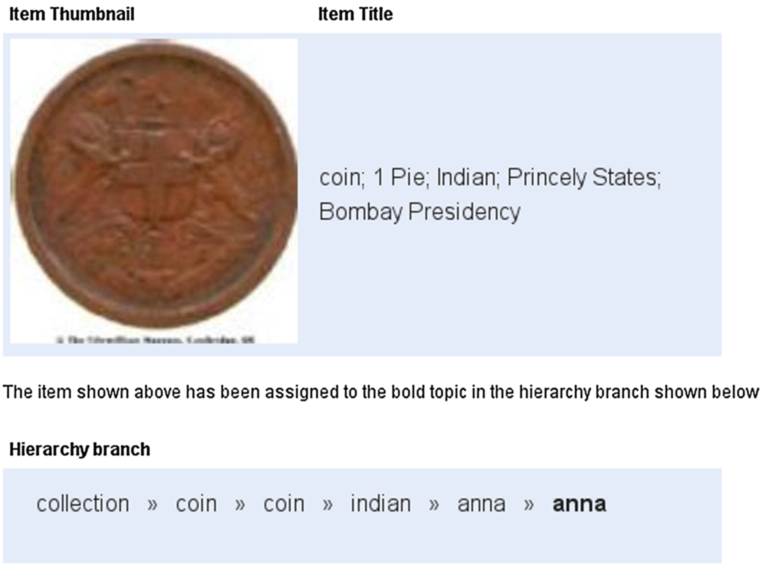
5.2 Item placement task

After exploring one of the hierarchies the participants moved to the second task in which they were asked to judge how well items were placed in the hierarchies. The aim of this task was to evaluate the hierarchies with respect to their ability of providing "context" information to the user (activity #2), by showing the user where in the hierarchy the current item is located and what topics it is related to.

5.2.1 Setup

A pool of 40 items was randomly selected from the 8,179 item test-set and, for each item, the branches leading from the root to the item were generated for each hierarchy. This created a set of 240 item–branch pairs from which participants were shown 10 randomly selected pairs. The sampling took into account the number of existing evaluations for each pair to ensure an even distribution. Due to limitations in the software the sampling did not take into account which hierarchy the participants had seen in the first task. This introduced a potential bias, as exposure to the hierarchy in the overviewing task might influence the item-placement judgement. However, analysis shows that there is no significant difference in the distributions of which hierarchy the participants previously used, across the item–branch pairs, negating the potential bias. For each item–branch pair, the participant was shown the branch from the root to the item, the item's title, keywords, and thumbnail (Fig. 8), and the following two questions:

1. *How well structured is the hierarchy branch?* (7-point semantic differential; *structured–unstructured*)
2. *Is this branch a suitable place for the item?* (7-point semantic differential; *suitable–not suitable*)



Item Thumbnail

Item Title

coin; 1 Pie; Indian; Princely States; Bombay Presidency

The item shown above has been assigned to the bold topic in the hierarchy branch shown below

Hierarchy branch

collection » coin » coin » indian » anna » **anna**

Fig. 8 Sample item with thumbnail, keywords, title, and the hierarchy branch as used in the item placement part of the first experiment

Table 4 Item placement task results

Hierarchy	1. Well-structured			2. Placed		
DBPedia	−2	0	2	−3	−2	0
LCSH	0	2	2	−2	0	2
LDA	−2	−1	2	−2	1	2
WikiFreq	**0**	**2**	**3**	**0**	**2**	**2**
WikiTax	0	2	2	−3	−1	1
WN Domains	−1	1	2	−3	−2	1

Numbers are on a seven-point scale from −3 (negative statement) to +3 (positive statement) and are shown 1st quartile/median/3rd quartile, with best numbers in bold

5.2.2 Results

The results (Table 4) clearly show that WikiFreq outperforms all other hierarchies, both on placement of items in the hierarchy (Wilcoxon signed-rank tests $p < 0.001$) and also on the structure of the hierarchy (Wilcoxon signed-rank test $p < 0.05$ against WikiTax & LCSH, $p < 0.001$ for all other hierarchies). The data-driven approach guarantees that the labels for the topics are directly linked to the item, and the arrangement of the topics in the hierarchy based on frequency ensures that the path to the item is seen as sensible. That this is an advantage of the data-driven approaches in general is confirmed by the LDA results, which outperform the WikiTax, DBPedia, and WN Domains hierarchies on item placement (Wilcoxon signed-rank tests $p < 0.001$), even though the structure scores are not good. While the lower scores are in line with Stoica et al. (2007), the good item placement scores indicate that the problem with LDA is likely related to the topic labelling and not to the structure in itself and scores could be improved by using a state-of-the-art labelling technique (Treeratpituk and Callan 2006; Lau et al. 2011).

The results show that the item placement algorithms for DBPedia, WikiTax, and WN Domains need to be improved, even if the branches themselves are relatively well structured. The mappings into LCSH are of variable quality, but do outperform the mappings of the other three manually curated hierarchies. One potential explanation for this is that the LCSH terms are more closely aligned with the language used in the data-set, ensuring that the mappings are better.

6 WikiMerge: a new hierarchy generation algorithm

The results in the previous section show that the data-driven approaches outperform the manual hierarchies on item placement quality, while the structure of the manually created hierarchies is judged as better overall. This is in line with our previous results where the data-driven approaches created more cohesive topics, but the manual hierarchies had better parent–child relationships. Based on this, a novel hierarchy was created that merged the data-driven WikiFreq approach for generating the leaf topics with a manually created hierarchy for the main structure.

Of the manually created hierarchies that had been tested, the results showed that DBPedia, WN Domains, and WikiTax were potential candidates. WN Domains gave the best overview, but in the previous experiment the parent–child relations were judged to be worse than in WikiTax. DBPedia did not give as good an overview, but the relationships were judged to be better than WN Domains, although still outperformed by WikiTax. WikiTax had the additional advantage that the Wikipedia language was likely to be more understandable for users (Milne et al. 2007) and with a few simple pruning rules the overviewing capabilities could also be improved. Based on this the decision was taken to use the WikiTax hierarchy and the resulting WikiMerge algorithm works as follows:

1. WikiFreq (Sect. 4.2) is used to link each item to Wikipedia articles $a_1 \ldots .a_n$, but only the link to a_n, the most specific article, is retained and the other links discarded.
2. The a_n articles, which represent the new hierarchy's leaf topics, are linked to their parent WikiTax topics based on the Wikipedia categories the articles belong to.
3. The resulting hierarchy is pruned removing all WikiTax topics that do not have a WikiFreq child or have only one child topic, reducing the WikiTax part of the hierarchy to the minimal hierarchy needed to structure the leaf topics.
4. The top-level topics in the combined hierarchy are then linked to their respective Wikipedia root node. This is done to correct an issue with the WikiTax construction method in Ponzetto and Strube (2011) that creates more root nodes than the 24 in Wikipedia.

The resulting WikiMerge hierarchy has WikiFreq topics as its leaves and WikiTax topics as its interior and root nodes and should thus merge its sources' strengths. The coverage for WikiMerge is the same as for WikiFreq (12.5 %) and after pruning to the shared 8,179 evaluation collection it has 20 root nodes, a total of 378 topics with a median depth of 3 (maximum 8) and median number of children of 2 (maximum 14). As the WikiMerge hierarchy is a combination of the WikiTax and WikiFreq hierarchies, its performance on the tasks in experiment 1 will mirror those of WikiFreq for the item placement and WikiTax for the overviewing task. To test this, a second experiment was devised, comparing WikiMerge (Fig. 9) to two of the previously tested hierarchies.

- Agriculture
- Applied sciences
  - Industrial design
    - Prototype
      - *coin, Fals, Islamic, Arab-Byzantine, Two standing figures (Heraclius prototype)*
- Arts
- Belief
  - Alternate reality
    - Fiction
      - Literature
        - *First Meeting of Dante and Beatrice*
- ... 18 further topics
- Science
- Society
- Technology

Fig. 9 An extract from the WikiMerge (Sect. 6) hierarchy. The extract demonstrates the narrow and deeper structure of the hierarchy and also the combination of the different types of topic titles (Wikipedia root topics at the top, then WikiTax topics, and finally WikiFreq leaf topics). The hierarchy extract was generated using the same algorithm as for the WikiTax hierarchy (Fig. 1)

7 Evaluation of WikiMerge

The previous experiments used four user-evaluated metrics to compare the hierarchies and based on the results we created an hybrid hierarchy merging two of the most useful hierarchies. The experiment described in this section was designed to test whether the user-evaluated metrics provide a good heuristic for the hierarchies' usefulness and whether the new WikiMerge hierarchy provides a better user experience, which would further validate the user-evaluated metrics.

The experiment was structured in three parts: an initial set of background questions, then a comparative study to gather user preferences, and finally the task-based activity. As with the previous experiments we used our own on-line experiment support system and the hierarchy browsing interface introduced in the previous experiment. Participants were recruited using the same method as in the previous experiments via the University's staff and student Volunteers mailing list, again without any incentives offered. The volunteers pool is essentially the same as for the first experiment, thus it is possible that participants from the first experiment also participated in this experiment. No personally identifiable information was acquired in either experiment, thus the degree to which this occurred cannot be quantified. There was a three month interval between the first and second experiment, which is significantly longer than most test–retest intervals (Falleti et al. 2006), and the experiment tasks were different, thus participation in the first experiment is unlikely to influence results in the second.

A total of 64 participants completed the experiment, of which 56 specified that their main language was English and only their responses are used in the analysis. Of the participants 34 were female and 22 male. 14 were between 18 and 25, 37 between 26 and 55, and the remaining 5 older. The gender distribution is the same as in the first experiment. The age distribution is slightly, but not statistically significantly higher in this experiment. Thus neither gender nor age are likely to impact the results. To further investigate our experiment population we asked participants whether they were studying (21), employed (34), or unemployed (1).

The experiment was limited to testing three hierarchies due to the complexity of the individual parts and to ensure that the length of the experiment did not exceed a reasonable time. LCSH was chosen as an example of a manually created hierarchy, as it does not perform badly on any aspects in the previous experiment, unlike DBPedia or WN Domains (although it does not perform as well as the other manually created hierarchies on some of the aspects). WikiFreq was chosen as the best-performing, fully automatically created hierarchy. WikiMerge was chosen to determine whether the conclusions we drew from the earlier results (and therefore the motivation for WikiMerge) were correct.

7.1 Comparing hierarchies

The first part of the experiment was designed to investigate which hierarchy participants preferred, assessing on the hierarchy activities #1 (“providing an overview”) and #4 (“organising the collection”). To determine preference, the participants were shown the hierarchies next to each other, an approach that has been shown to be successful in IR evaluations (Carterette et al. 2008).

7.1.1 Setup

Participants were shown the three hierarchies next to each other and JavaScript was used to ensure that the hierarchy display covered the full height of the window. The order of the three hierarchies was randomly assigned to each participant and the three hierarchies were always labelled “A”, “B”, and “C” regardless of the displayed order to ensure that no ordering bias was introduced. Participants were instructed to spend a few minutes exploring the three hierarchies and then to scroll down and answer the questions listed below:

1. *For each of the three hierarchies shown above rate how understandable the individual headings are* (5-point semantic differential; *not at all understandable–very understandable*)
2. *For each of the three hierarchies shown above rate how well the headings are organised* (5-point semantic differential; *very badly organised–very well organised*)
3. *In general which of the three hierarchies do you prefer?* (choice A, B, C)
4. *If you were looking for a specific topic, which of the hierarchies would you prefer?* (choice A, B, C)
5. *If you were trying to re-find an item you had previously viewed, which of the three hierarchies would you prefer?* (choice A, B, C)
6. *Please briefly explain why you prefer the selected hierarchy* (free text)

7.1.2 Results

A Kruskal-Wallis test shows no significant influence of order on preference ($\chi^2 = 2.45, df = 5, p = 0.78$) or time spent answering the questions ($\chi^2 = 5.97, df = 5, p = 0.31$). The results in Table 5 show a clear preference for the WikiMerge hierarchy. Both the “understandable” and “organisation” scores are significantly higher than for the other two hierarchies (Wilcoxon rank-sum test $p < 0.001$). This is confirmed by the “preference” selection, where two-thirds of the participants prefer the WikiMerge hierarchy, with the remaining third split relatively evenly between LCSH and WikiFreq.

When asked about their preference for finding a specific topic, five of the participants who initially selected WikiMerge selected LCSH instead, but all five (plus an additional 4) switched back to WikiMerge for the re-finding question. The most likely explanation for

Table 5 Preference experiment results including WikiMerge

	LCSH			WikiFreq			WikiMerge		
1. Understandable	−1	0	1	−1	0	1	**1**	**1.5**	**2**
2. Organisation	−1	0	1	−1	0	1	**0**	**1**	**1**
3. Preference	12			8			36		
4. Topic	12			10			34		
5. Re-find	8			8			40		

The first two questions are on a five-point scale from −2 (negative statement) to +2 (positive statement) and are shown 1st quartile/median/3rd quartile, with best numbers in bold. The other three questions are the number of participants who selected that hierarchy

this is that for those five participants LCSH's flat, alphabetical structure means that they believed that finding a specific topic would require only scrolling through the list and not exploring the individual branches. However, it is clear that this same breadth at the top level means that for navigation there are less landmarks to help the participants remember where the topic had been located, making re-finding more difficult.

7.2 Exploration task

The results so far have shown that the WikiMerge hierarchy is preferred over the other two hierarchies. The goal of the second part of the experiment was to quantify how useful the hierarchies would be in a task context and whether the preference for WikiMerge would also lead to higher task performance.

7.2.1 Setup

The exploration task was split over two pages. On the first page the users were shown the task instructions and below an interface to explore the hierarchy, view the items, and save those items that they felt were relevant for their task. The participants were instructed to complete the task and then move on to the next page to answer the following questions:

1. *How easy was it to navigate the hierarchy?* (5-point semantic differential; *very difficult–very easy*)
2. *How easy was it to find the items you selected?* (5-point semantic differential; *very difficult–very easy*)
3. *How satisfied are you with the items you found?* (5-point semantic differential; *very unsatisfied–very satisfied*)
4. *How successful do you feel you were in completing the task?* (5-point semantic differential; *very unsuccessful–very successful*)
5. *How useful would you find this kind of interface in practice?* (5-point semantic differential; *not very useful–very useful*)
6. *What did you like about the hierarchy?* (free text)
7. *What did you dislike about the hierarchy?* (free text)

The task instructions were derived from Skov and Ingwersen (2008)'s "simulated leisure task" and consisted of a thumbnail image that acted as a stimulus and a short paragraph explaining the task context. Briefly summarised the three tasks descriptions were:

Table 6 Exploration task aggregate results by hierarchy

	LCSH			WikiFreq			WikiMerge		
1. Navigate	−1	−0.5	0	−1	0	1	−1	−0.5	1
2. Find items	−1	−1	0	−1	0	1	−1	0	0
3. Satisfied	−1.25	−0.5	1	−1.75	−1	0	−1	−1	0.75
4. Success	−2	−1	1	−1.75	−1	0	−1	−1	0.75
5. Useful	−2	−1	0	−1	0	1	−1.75	−1	0.75
Items	1	3	11	1	6	11	3.75	8.5	11
Time(s)	249	351	454	198	279	416	240	368	502

Results to the questions and number of items found. The responses for the first five questions are on a five-point scale from −2 (negative statement) to +2 (positive statement). All values are shown 1st quartile/median/3rd quartile

- *Calendar*: Find 11 items to combine with the thumbnail image to create a calendar.
- *Coin*: Find a few items like the thumbnail image of a coin that you found while walking.
- *Presentation*: For a short presentation find a few items that fit thematically to the thumbnail image.

The stimulus image and descriptions were chosen based on the items available in the collection, in order to ensure that for each task a large pool of potentially relevant items was available for the participants to find and choose from. Also the open-ended nature of the three task contexts ensures that browsing the hierarchy is a realistic approach to completing the tasks. While all three tasks are open-ended, by design they are very different in nature, as we wished to evaluate the hierarchies' usefulness across a range of tasks.

After having seen all three hierarchies in the previous task, for this task the same participants were randomly assigned one hierarchy and one task, with the sampling ensuring that each pair was evaluated the same number of times. With a total of 56 participants, this results in 6 or 7 participants per task–hierarchy combination. While not as many as initially planned, the numbers are sufficient to allow a statistical comparison.

7.2.2 *Results*

The aggregate results for each hierarchy are shown in Table 6. Kruskal–Wallis tests were used to investigate if the hierarchy and task pairs had any statistical influence on the user-provided measures. None of the tests were significant, indicating that none of the tasks favoured one of the hierarchies. The number of items found is statistically significantly different based on hierarchy and task ($\chi^2 = 31.54, df = 8, p < 0.001$), however a decomposition shows that only the task has a significant influence ($\chi^2 = 28.67, df = 2$, $p < 0.001$), while the hierarchy used has no influence. The cause for this influence is that the tasks did not specify that the same number of items should be found. Thus the significant difference by task has little intrinsic importance.

The first thing that is clear from the results is that across all hierarchies the results are very poor. Two possible explanations for this are that either the hierarchies are not very good or that issues with the collection are impacting the results. The second possibility is likelier, because although almost all the items the users found were relevant to their respective tasks (377 out of 400 items—94.25 %), the participant's evaluation of how satisfied they were with the items was low. This may be largely due to the limited amount of meta-data available for each item, particularly the small size of the thumbnail images. This lack of information led to a feeling of dissatisfaction with the items, which in turn influences the “success” evaluation, as evidenced by the strong correlation between the “satisfied” and “success” ratings (Spearman’s $\rho = 0.849, p < 0.001$).

The results also show that the participants are able to work around the limitations of the individual hierarchies and to successfully complete all the tasks with any of the hierarchies. Thus to support more open-ended, exploratory interactions with a collection, any hierarchy can support the user in their task. The question of which hierarchy to choose thus transforms into the question of which hierarchy the potential user is most likely to be comfortable with. For systems targeted at information professionals, a known hierarchy such as LCSH might be preferable, while the more common language used in the Wikipedia-derived hierarchies makes those hierarchies more accessible to non-specialist users. The wide coverage in Wikipedia and the generic, data-driven nature of the algorithms also means that they can be applied to any collection that is topically covered by Wikipedia.

While not statistically significant, an interesting result is that in the “coin” task, seven participants failed to find any items, indicating that the task was potentially harder than the other two tasks, which were successfully completed by all participants. More importantly the failures were not evenly distributed across the three hierarchies. For both the WikiFreq and LCSH hierarchies, three participants failed to find any items, while for WikiMerge it was only one. This could potentially indicate that there is some aspect of the WikiMerge hierarchy that helped avoid participants get completely lost. However, further experiments are needed to investigate this.

7.3 Qualitative results

In both parts of the second experiment the participants were asked to answer a set of qualitative questions. “Why did you prefer this hierarchy?” in the first part and “What did you like about this hierarchy?” and “What did you dislike about this hierarchy?” in the second part. An inductive approach was taken to categorise the comments, where each comment could receive multiple categories. Table 7 shows the results for categories that were mentioned at least twice in the answers.

The qualitative responses are in line with the quantitative results, with many positive comments for WikiMerge in the qualitative question asked as part of the hierarchy comparison, but a more even spread of positive and negative comments after completing the exploration task. It is clear that what people like about the WikiMerge hierarchy is that there is a clear drill-down structure to the hierarchy, that the concepts have labels which the users understand and that the top-level concepts are clearly organised. The clear drill-down and understandable labels are also mentioned for WikiFreq, indicating that Wikipedia is generally a useful intermediary. However, the fact that no participant mentioned a clear top-level for WikiFreq indicates that the purely data-driven approaches create top-level structures that initially seem unclear and potentially unstructured. For the LCSH-based hierarchy people commented on the alphabetic ordering being something they appreciated.

An interesting aspect in the qualitative responses gathered after the exploration task is that getting lost in the hierarchy and feeling disoriented is a problem for both LCSH and WikiMerge. This is mirrored by the fact that the fraction of participants who say they liked the clear drill-down aspect of the hierarchies is lower, particularly noticeable with WikiMerge. WikiFreq performs slightly better, potentially because the pure data-driven approach ensures that the topic labels are closer to the task and thus easier to drill down into. This can also be seen in the quantitative “navigate” and “find items” questions, where WikiFreq has the highest scores.

Table 7 Main categories for the three qualitative questions

	LC (12)	WF (8)	WM (36)
Preference			
Clear drill-down	3 (.25)	4 (.5)	**19 (.53)**
Understandable labels	2 (.16)	2 (.25)	**14 (.38)**
Clear top-level	0	0	**11 (.31)**
Size	3 (.25)	2 (.25)	5 (.14)
Alphabetic	**4 (.33)**	0	1 (.03)
Mapping	–	–	–
Disorientation	–	–	–
	LC (20)	WF (18)	WM (18)
Like			
Clear drill-down	3 (.15)	**5 (.27)**	4 (.22)
Understandable labels	0	**2 (.11)**	1 (0.05)
Clear top-level	–	–	–
Size	2 (.1)	0	0
Alphabetic	1 (.05)	1 (.06)	1 (.06)
Mapping	0	1 (.06)	0
Disorientation	–	–	–
	LC (20)	WF (18)	WM (18)
Dislike			
Clear drill-down	–	–	–
Understandable labels	0	**2 (.11)**	0
Clear top-level	–	–	–
Size	1 (.05)	0	0
Alphabetic	**1 (.05**)	0	0
Mapping	**4 (.2)**	1 (.55)	0
Disorientation	5 (.25)	3 (.16)	**6 (.33)**

Each of the participants' answer can be assigned to multiple categories. Clear drill-down: the users mentioned that at each level the choice of child topics made sense with respect to the parent topic; Understandable labels: the users mentioned that the labels were clearly understandable; Clear top-level: the users mentioned that the root topics were distinct from each other; Size: the users mentioned that the amount of information shown was good; Alphabetic: the users mentioned that the alphabetical ordering was useful; Mapping: the item to topic mappings were judged to be good/bad; Disorientation: essentially the opposite of "clear drill-down" in that the users did not know where to go next or felt that the next level did not fit with the parent. Bold numbers indicate the hierarchy with the most mentions of the respective categories

8 Final discussion

Considering all the results we can conclude that the four proposed user-evaluated metrics (topic cohesion, parent–child relationships, item placement, and overview) are well suited to evaluating a hierarchy and give a good prediction of what hierarchies will have higher user preference. The user-evaluated metrics are also in line with the characteristics that participants mentioned in their qualitative responses, namely a clearly understandable structure that supports drill-down into the collection (tested by item placement and parent–child relationship questions), good mappings of items into the hierarchy (topic cohesion and item placement), and a clear set of top-level topics (overview). The qualitative results also support the use of Wikipedia as an intermediary, particularly for non-expert users, as it provides understandable labels.

A further advantage of the four user-evaluated metrics is that they enable a clearer distinction between the hierarchies. The larger, task-based evaluation did not show any significant difference between the three hierarchies we tested. This is most likely due to the participants being able to work around the hierarchies' shortcomings and successfully complete the tasks using all three tested hierarchies. While the context is more realistic, the lack of distinguishable results means that the explanatory power of the experiment is limited.

An interesting question that is raised by the analysis is whether items should be included in the evaluation or not since our results suggest that this can affect how users perceive hierarchies. Consider two questions which address the same structural aspect: the "Organised" and "Well-structured" questions in the hierarchy comparison experiments (Tables 3, 4 respectively). In those two questions the hierarchies that perform well on the item placement question (LDA, WikiFreq) have higher "well-structured" ratings than for the "organised" question. The opposite is true for those hierarchies that do not perform well on the item placement (DBPedia, WN Domains). Some previous work on the evaluation of hierarchies for navigation chose not to include items (Stoica et al. 2007) while other researchers did include them (Milne et al. 2007; Sanderson and Croft 1999). It is possible that the preference for a manually created hierarchy that was observed when items were not included (Stoica et al. 2007) and the success of automatic methods when they were (Milne et al. 2007; Sanderson and Croft 1999) was at least in part due to this choice.

This aspect seems to argue for the inclusion, however in the task-based evaluation we see a strong correlation of the participants' satisfaction with the items and their self-evaluation of task success, which is not in line with an objective success assessment where 87.5 % of the participants are judged as being successful. While this validates that participants were engaging with the task and not treating it as an abstract exercise, it does mean that for a task-based evaluation it is important to ensure that the items used are not only relevant, but also engaging. This highlights a further advantage of the four user-evaluated metrics, as with them items are only shown when absolutely necessary, limiting the impact of their engagingness, while still providing reliable judgements that enable the comparison of the tested hierarchies.

While we tested the algorithms on a Cultural Heritage collection, none of the algorithms are domain-specific or use domain-specific inputs and are thus in theory generalisable to any domain. The only constraint is that where the algorithms use external resources (LCSH, Wikipedia, WordNet Domains,…), these have to cover the target collection's topics. While determining the amount of coverage that the external resources provide is beyond the scope of this research, it is likely that the more specific the content of a collection to process, the less probable it is that the generic resources provide good coverage.

The results from the individual user-evaluated metrics are grounded in the type of participants that we recruited. As we were primarily interested in the usefulness of hierarchies for the novice user, the participant population were generally non-experts. Thus the results and preferences are only directly applicable to similar user groups and it is unclear to what degree the results generalise to other user groups, such as expert users. However, the ease with which the individual aspects can be tested means that a set of potential hierarchies for a given collection can quickly and easily be evaluated with participants drawn from the target user group. This means that the choice of the "most useful" hierarchy can be chosen based on the specific target collection and target user group.

9 Summary

In this paper we have experimented with different approaches for automatically generating subject hierarchies for a digital library collection and then evaluating these. These hierarchies could support more exploratory forms of information seeking. Various existing lexical resources, such as LCSH, WordNet Domains and Wikipedia Categories, were used to derive concepts in the hierarchy and provide a parent–child structure. In addition purely data-driven approaches were also used, including LDA and a method based on identifying relevant Wikipedia articles (WikiFreq).

Various novel techniques were used to evaluate the hierarchies, assessing different aspects of the hierarchies, including the gathering of user preferences and a task-based study where users had to check item placement. The users tended to prefer small manual hierarchies like WordNet Domains or DBPedia ontology for exploration, and one of the automatically produced hierarchy (WikiFreq) for item placement. From the qualitative responses we gathered the following three core characteristics emerge that a good hierarchy must have: a clearly understandable structure that supports drill-down into the collection, good mappings of items into the hierarchy, and a clear set of top-level topics.

Those results lead to the development of WikiMerge, a new algorithm for generating hierarchies that combines a data-driven bottom-up approach (WikiFreq) that provides good topic–item cohesion and mappings with a Wikipedia-based taxonomy (WikiTax) to impose a sensible hierarchical structure. In addition the taxonomy is pruned to only keep the branches which are necessary to cover the concepts present in the collection, resulting in a domain-specific hierarchy. The broad coverage of Wikipedia means that the algorithm can be applied to a wide variety of domains. Due to its narrow top-level structure the WikiMerge hierarchy could also easily be transformed into a set of hierarchical facets as used in Hearst (2006a).

The WikiMerge hierarchy was compared head-to-head with the manual hierarchy that performed well in the first experiments (LCSH), and with the best of the automatic hierarchies (WikiFreq). Participants clearly preferred the new hierarchy over the other two hierarchies. However, in a task-based evaluation we found no significant differences in performance between the three hierarchies. Participants succeeded in completing the three tasks that they were given with all hierarchies. This leads us to the conclusion that as long as a hierarchy provides coverage over the collection, the users can work around the limitations of the hierarchy to solve their task.

Evaluating hierarchies is a complex task, particularly when focusing on the interaction of the user with the hierarchy. This work presents a first step towards a set of user-focused evaluation methods that can easily and quickly be applied to any number of hierarchies or target user groups.

In future work we intend to extend this work in three directions. First, we intend to investigate how stable the results are when the hierarchy algorithms are applied to collections in other domains and tested with different user groups (expert vs. non-expert). For example, the medical domain is promising since hierarchies such as Medical Subject

Headings (MeSH) are available while both expert and non-expert users (i.e. medical practitioners and patients respectively) are interested in finding information. Second, we intend to further investigate whether it is possible to create task-based evaluations that clearly differentiate the properties of hierarchies. Finally, the way in which a hierarchy is presented to the user is likely to impact its attractiveness and usefulness to the user and we are planning to evaluate a number of visualisation interfaces for navigating hierarchies.

Acknowledgments The research leading to these results was supported by the PATHS project (http://paths-project.eu) funded by the European Community's Seventh Framework Programme (FP7/2007-2013) under Grant Agreement No. 270082.

References

Anick, P., & Tipirneni, S. (1999) The paraphrase search assistant: Terminological feedback for iterative information seeking. In *Proceedings of the 22nd annual international ACM SIGIR conference on research and development in information retrieval, ACM* (pp. 153–159).

Atserias, J., Villarejo, L., Rigau, G., Agirre, E., Carroll, J., Magnini, B. et al. (2004). The meaning multilingual central repository. In *Proceedings of GWC* (pp. 23–30).

Auer, S., Bizer, C., Kobilarov, G., Lehmann, J., Cyganiak, R., & Ives, Z. (2007). Dbpedia: A nucleus for a web of open data. In *The Semantic Web* (pp. 722–735).

Azzopardi, L., Girolami, M., & Van Rijsbergen, C. (2004). Topic based language models for ad hoc information retrieval. In *Neural networks, 2004. Proceedings. 2004 IEEE international joint conference on, IEEE* (Vol. 4, pp. 3281–3286).

Blei, D. M., Griffiths, T., Jordan, M., & Tenenbaum, J. (2003). Hierarchical topic models and the nested chinese restaurant process. In NIPS. http://books.nips.cc/papers/files/nips16/NIPS2003_AA03.pdf.

Borlund, P., & Ingwersen, P. (1997). The development of a method for the evaluation of interactive information retrieval systems. *Journal of documentation*, *53*(3), 225–250.

Brewster, C., Alani, H., Dasmahapatra, S., & Wilks, Y. (2004). Data driven ontology evaluation. In *Proceedings of international conference on language resources and evaluation.*

Carterette, B., Bennett, P. N., Chickering, D. M., & Dumais, S. T. (2008). Here or there: Preference judgements for relevance. In C. Macdonald, I. Ounis, V. Plachouras, I. Ruthven, & R. W. White (Eds.), *Proceedings of the IR research, 30th European conference on advances in information retrieval (ECIR'08)* (pp. 16–27). Berlin, Heidelberg: Springer.

Chang, J., Boyd-Graber, J., Wang, C., Gerrish, S., & Blei, D. M. (2009). Reading tea leaves: How humans interpret topic models. In *NIPS.*

Chen, M., Hearst, M., Hong, J., & Lin, J. (1999). Cha-cha: A system for organizing intranet search results. In *Proceedings of the 2nd conference on USENIX symposium on internet technologies and systems* (pp. 11–14).

Falleti, M. G., Maruff, P., Collie, A., & Darby, D. G. (2006). Practice effects associated with the repeated assessment of cognitive function using the cogstate battery at 10-minute, one week and one month test–retest intervals. *Journal of Clinical and Experimental Neuropsychology*, *28*(7), 1095–1112.

Fellbaum, C. (1998). *WordNet: An electronic database*. Cambridge, MA: MIT Press.

Fernando, S., Hall, M., Agirre, E., Soroa, A., Clough, P., & Stevenson, M. (2012). Comparing taxonomies for organising collections of documents. In *Proceedings of COLING 2012, The COLING 2012 Organizing Committee, Mumbai, India* (pp. 879–894). http://www.aclweb.org/anthology/C12-1054.

Gómez-Pérez, A. (1996). Towards a framework to verify knowledge sharing technology. *Expert Systems with Applications*, *11*(4), 519–529.

Hall, M. M., & Toms, E. (2013). Building a common framework for iir evaluation. In *CLEF 2013—Information access evaluation. Multilinguality, multimodality, and visualization* (pp. 17–28). doi:10.1007/978-3-642-40802-1_3.

Hearst, M. (2006a). Clustering versus faceted categories for information exploration. *Communications of the ACM*, *49*(4), 59–61.

Hearst, M. (2006b). Design recommendations for hierarchical faceted search interfaces. In *Proceedings of the 29th annual international ACM SIGIR conference on research and development in information retrieval (SIGIR'06) workshop on faceted search.*

Hoffart, J., Suchanek, F., Berberich, K., Lewis-Kelham, E., De Melo, G., & Weikum, G. (2011). Yago2: Exploring and querying world knowledge in time, space, context, and many languages. In *Proceedings of the 20th international conference companion on World Wide Web, ACM* (pp. 229–232).

Hornbæk, K., & Hertzum, M. (2011). The notion of overview in information visualization. *International Journal of Human-Computer Studies, 69*(7–8), 509–525. doi: 10.1016/j.ijhcs.2011.02.007. http://www.sciencedirect.com/science/article/B6WGR-529V18J-1/2/95a091a9a1a8d5423cd3fbdbd6ff5fc2.

Horvat, M., Grbin, A., & Gledec, G. (2012) Wntags: A web-based tool for image labeling and retrieval with lexical ontologies. In: M. Graña, C. Toro, J. Posada, R. J. Howlett & L. C. Jain (Eds.), *Frontiers in artificial intelligence and applications* (Vol. 243, pp. 585–594). KES, IOS Press.

Jörgensen, C. (2004). Unlocking the museum: A manifesto. *Journal of the American Society for Information Science and Technology*, *55*(5), 462–464. doi:10.1002/asi.10396.

Kelly, D., & Sugimoto, C. (2013). A systematic review of interactive information retrieval evaluation studies, 1967–2006. *JASIST*, *64*(4), 745–770.

Lau, J., Grieser, K., Newman, D., & Baldwin, T. (2011). Automatic labelling of topic models. In *Proceedings of the 49th annual meeting on association for computational linguistics* (pp. 1536–1545).

Lawrie, D., Croft, W., & Rosenberg, A. (2001). Finding topic words for hierarchical summarization. In *Proceedings of the 24th annual international ACM SIGIR conference on research and development in information retrieval, ACM* (pp. 349–357).

Liu, X., Song, Y., Liu, S., & Wang, H. (2012). Automatic taxonomy construction from keywords. In *Proceedings of the 18th ACM SIGKDD international conference on knowledge discovery and data mining, ACM, New York, NY, USA, KDD '12* (pp. 1433–1441). doi:10.1145/2339530.2339754.

Maedche, A., & Staab, S. (2002). Measuring similarity between ontologies. In *Knowledge engineering and knowledge management: Ontologies and the semantic web* (pp. 15–21).

Magnini, B., & Cavaglia, G. (2000). Integrating subject field codes into wordnet. In *Proceedings of LREC-2000, second international conference on language resources and evaluation* (pp. 1413–1418).

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

Markkula, M., & Sormunen, E. (2000). End-user searching challenges indexing practices in the digital newspaper photo archive. *Information Retrieval*, *1*(4), 259–285.

Milne, D., & Witten, I. H. (2008). Learning to link with Wikipedia. In *Proceedings of the 17th ACM conference on information and knowledge management* (pp. 509–518).

Milne, D. N., Witten, I. H., & Nichols, D. M. (2007). A knowledge-based search engine powered by Wikipedia. In *Proceedings of the sixteenth ACM conference on conference on information and knowledge management, ACM* (pp. 445–454).

Navigli, R., Velardi, P., & Gangemi, A. (2003). Ontology learning and its application to automated terminology translation. *Intelligent Systems, IEEE*, *18*(1), 22–31.

Nevill-Manning, C., Witten, I., & Paynter, G. (1999). Lexically-generated subject hierarchies for browsing large collections. *International Journal on Digital Libraries*, *2*(2), 111–123.

Padró, L., Reese, S., Agirre, E., & Soroa, A. (2010). Semantic services in freeling 2.1: Wordnet and ukb. In P. Bhattacharyya, C. Fellbaum, & P. Vossen (Eds.), *Principles, construction, and application of multilingual Wordnets, global Wordnet conference 2010* (pp. 99–105). Mumbai, India: Narosa Publishing House.

Pirolli, P. (2009). Powers of 10: Modeling complex information-seeking systems at multiple scales. *Computer*, *42*(3), 33–40. doi:10.1109/MC.2009.94.

Pirolli, P., Schank, P., Hearst, M., & Diehl, C. (1996). 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, ACM* (pp. 213–220).

Ponzetto, S., & Strube, M. (2011). Taxonomy induction based on a collaboratively built knowledge repository. *Artificial Intelligence*, *175*(9–10), 1737–1756.

Pratt, W., Hearst, M., & Fagan, L. (1999). A knowledge-based approach to organizing retrieved documents. In *Proceedings of th 16th annual conference on artificial intelligence (AAAI 99).*

Rao, R., Pedersen, J. O., Hearst, M. A., Mackinlay, J. D., Card, S. K., Masinter, L., et al. (1995). Rich interaction in the digital library. *Communications of the ACM*, *38*(4), 29–39. doi:10.1145/205323.205326.

Rosenfeld, L., & Morville, P. (2002). *Information architecture for the World Wide Web: Designing large-scale web sites*. Sebastopol: O'Reilly Media, Incorporated.

Sanderson, M., & Croft, B. (1999). Deriving concept hierarchies from text. In *Proceedings of the 22nd annual international ACM SIGIR conference on research and development in information retrieval, ACM* (pp. 206–213).

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

Singer, G., Norbisrath, U., & Lewandowski, D. (2012). Ordinary search engine users carrying out complex search tasks. *Journal of Information Science*, *39*(3), 346–358.

Skov, M., & Ingwersen, P. (2008). Exploring information seeking behaviour in a digital museum context. In *Proceedings of the second international symposium on Information interaction in context, ACM* (pp. 110–115).

Stoica, E., Hearst, M., & Richardson, M. (2007). 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)* (pp. 244–251).

Tang, L., Zhang, J., & Liu, H. (2006). Acclimatizing taxonomic semantics for hierarchical content classification. In *Proceedings of the 12th ACM SIGKDD international conference on knowledge discovery and data mining, ACM, New York, NY, USA, KDD '06* (pp. 384–393). doi:10.1145/1150402.1150446.

Toms, E. G., Villa, R., & McCay-Peet, L. (2013). How is a search system used in work task completion? *Journal of Information Science*, *39*(1), 15–25.

Treeratpituk, P., & Callan, J. (2006). Automatically labeling hierarchical clusters. In *Proceedings of the 2006 international conference on Digital Government Research, Digital Government Society of North America, dg.o '06* (pp. 167–176). doi:10.1145/1146598.1146650.

Wang, Z., Khoo, C. S., & Chaudhry, A. S. (2014). Evaluation of the navigation effectiveness of an organizational taxonomy built on a general classification scheme and domain thesauri. *Journal of the Association for Information Science and Technology*,. doi:10.1002/asi.23017.

Wei, X., & Croft, W. (2006) LDA-based document models for ad-hoc retrieval. In *Proceedings of the 29th annual international ACM SIGIR conference on research and development in information retrieval, ACM* (pp. 178–185).

White, R. W., Kules, B., Drucker, S. M., & Schraefel, M. (2006). Introduction. *Communications of the ACM*, *49*(4), 36–39. doi:10.1145/1121949.1121978.

Yakel, E., Shaw, S., & Reynolds, P. (2007). Creating the next generation of archival finding aids. *D-Lib Magazine*, *13*(5/6). doi:10.1045/may2007-yakel.

Yu, J., Thom, J., & Tam, A. (2007). Ontology evaluation using Wikipedia categories for browsing. In *Proceedings of the sixteenth ACM conference on conference on information and knowledge management, ACM* (pp. 223–232).