Comparing taxonomies for organising collections of documents

Samuel Fernando[1] *Mark Hall*[1,2] *Eneko Agirre*[3]
Aitor Soroa[3] *Paul Clough*[2] *Mark Stevenson*[1]
(1) Department of Computer Science, Regent Court, 211 Portobello, Sheffield, S1 4DP, UK
(2) Information School, Regent Court, 211 Portobello, Sheffield, S1 4DP, UK
(3) Universidad del País Vasco, Barrio Sarriena s/n, 48940 Leioa, Bizkaia
{s.fernando,m.mhall,p.d.clough,m.stevenson}@sheffield.ac.uk
{e.agirre, a.soroa}@ehu.es

ABSTRACT
There is a demand for taxonomies to organise large collections of documents into categories for browsing and exploration. This paper examines four existing taxonomies that have been manually created, along with two methods for deriving taxonomies automatically from data items. We use these taxonomies to organise items from a large online cultural heritage collection. We then present two human evaluations of the taxonomies. The first measures the cohesion of the taxonomies to determine how well they group together similar items under the same concept node. The second analyses the concept relations in the taxonomies. The results show that the manual taxonomies have high quality well defined relations. However the novel automatic method is found to generate very high cohesion.

TITLE AND ABSTRACT IN BASQUE

Dokumentu bildumak antolatzeko taxonomien arteko alderaketa

Dokumentu bildumak kategorietan sailkatzea oso erabilgarria da, dokumentuak arakatzeko eta aztertzeko aukera berriak eskaintzen duen heinean. Hori horrela izanik, dokumentu bilduma handiak sailkatzeko taxonomien behar handia dago. Artikulu honetan eskuz sortutako lau taxonomia aztertzen dira, taxonomiak automatikoki sortzen dituzten bi metodorekin batera. Taxonomia hauek ondare kulturaleko bilduma handiak antolatzeko erabili ditugu. Taxonomien ebaluazioa egin dugu galdetegietan oinarritutako bi metodo erabiliaz. Lehenbizikoak taxonomiaren kohesioa neurtzen du, hau da, antzeko itemak kontzeptu beraren azpian zein ondo taldekatzen diren. Bigarrenak taxonomiako kontzeptuen arteko erlazioak aztertzen ditu. Emaitzek erakusten dute eskuzko taxonomien erlazioen kalitatea, baina metodo automatiko berri batek lortzen du kohesio handiena.

KEYWORDS: Semantic network, taxonomy, hierarchy, Wikipedia, WordNet, ontology.

KEYWORDS IN BASQUE: Sare semantiko, taxonomia, hierarkia, Wikipedia, WordNet, Ontologia.

Proceedings of COLING 2012: Technical Papers, pages 879–894,
COLING 2012, Mumbai, December 2012.

1 Introduction

With increasingly large sets of diverse collections of documents available online a key challenge is organising and presenting these items effectively for information access. To enable the navigation and exploration of collections, content providers typically provide users with free-text search functionalities, along with some form of browsable subject categories or taxonomy, also useful in organising documents. Providing multiple mechanisms for accessing documents enables users to conduct various modes of information seeking activity, from locating specific documents to more exploratory forms of searching and browsing behaviour (Hearst, 2009; Marchionini, 2006; Wilson et al., 2010).

In this paper we focus on evaluating different taxonomies that could be used to organise and navigate content from Europeana[1], an online cultural heritage collection. This collection comprises many subcollections taken from different providers, and thus contains a very diverse set of cultural heritage *items*[2]. Some of the subcollections are linked to bespoke taxonomies; however, many are not. This therefore represents a very challenging dataset to organise in a consistent and uniform manner. We focus on two main approaches for organising content: the first is to map items from Europeana onto existing manually-created taxonomies; the second is to use data-driven approaches to automatically derive taxonomies from the collection. This requires being able to successfully group items into categories and generate suitable category labels. A note on definitions: the term 'taxonomy' is used in this paper as a general term meaning a conceptual hierarchy. A taxonomy does not necessarily have to be a subsumption hierarchy (where each child concept is subsumed by its parent concept). Some of the taxonomies described here are subsumption hierarchies and some are not.

There are many different ways of evaluating taxonomies (Snow et al., 2004; Malaisé et al., 2006; Yi, 2008; Nikolova et al., 2010; Ponzetto and Strube, 2011). Here we focus on two approaches which capture different qualities of the taxonomies. The first evaluation measures the *cohesion* of the taxonomies; how well they group together similar items into the same concept node. The second analyses the *relationships* between concept nodes in the taxonomies and whether people can understand the concept labels. We believe this is the first time that such evaluations have been applied to such taxonomies over large collections of data items.

This paper provides three main contributions: (1) the systematic comparison of different taxonomies for organising a large cultural heritage collection; (2) a novel data-driven approach based on using Wikipedia article links as concept nodes in the taxonomy; and (3) the evaluation of cohesion and relationship type between concepts using an approach based on crowdsourcing. The rest of the paper comprises the following. Section 2 describes related work in this area. Section 3 gives a brief overview of the key data resources and tools referenced and Section 4 describes the taxonomies and item-to-resource mapping approaches used in the experiments. Section 5 describes the cohesion and relatedness experiments and the results obtained.

2 Related work

There are many category systems or taxonomies available, some of which are domain specific; others aimed at covering more general subjects. For example, one of the most popular and commonly used resources is the Library of Congress Subject Headings (LCSH)[3]. LCSH provides

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

[2] An item is defined here as a online record of a cultural heritage artifact (usually an image), together with associated metadata, such as title, subject, description etc.

[3] `http://www.loc.gov/aba/cataloging/subject/`

a controlled vocabulary of keywords (or subject headings), which are widely used in libraries to catalogue materials and facilitate information access. Similarly, in the medical domain MeSH[4], created and maintained by the National Library of Medicine, provides a controlled vocabulary of medical subject headings. In computational linguistics, WordNet (Fellbaum, 1998) is a commonly used lexical knowledge base that links concepts in various ways. WordNet has been expanded with WordNet domain labels which group together words from different syntactic categories and different senses (Bentivogli et al., 2004). These domain labels are organised into a hierarchical structure.

There is a body of previous work on automatically deriving taxonomies and relations from free text. Hearst (1992) was perhaps the earliest significant effort to derive hyponym-hypernym relations from free text using the now eponymous Hearst patterns which code common syntactic forms of the hyponymy pattern (e.g. 'Vehicles such as Cars'). These patterns were hand-crafted. More recently Snow et al. (2004) developed on this work by using an existing knowledge base to automatically derive lexico-syntactic patterns containing the hyponym-hypernym pairs.

An alternative to creating hierarchies of concepts from the pattern-based methods is to use statistical methods. Sanderson and Croft (1999) used an approach to automatically build a hierarchy of terms or concepts (nouns and noun phrases) based on term co-occurrences within a set of documents. To order the concepts a statistical relation called *subsumption* was used to determine which of a co-occurring pair of concepts was most likely to be the parent. Griffiths and Tenenbaum (2004) uses Bayesian methods and LDA (Latent Dirichlet Allocation) to derive topic clusters and create a topic hierarchy. Recent work has exploited the information in Wikipedia to create taxonomies. Ponzetto and Strube (2011) uses the category hierarchy along with lexical matching methods to create a taxonomy which compares well with manually created resources. DBpedia (Auer et al., 2007) also uses the Wikipedia category hierarchy but also additionally links in the articles into the hierarchy.

One key problem is how to evaluate taxonomies effectively. This is a complex problem since there are many aspects to consider. We can consider how users would rate various aspects of their experience using a questionnaire. However this subjective study can be misleading since people can underrate or overrate their experience. A more objective approach is to log user interactions and then infer from these how effective the taxonomy is i.e. time spent on a task, how much of the domain was covered and so on. There have been user studies of taxonomies which have used combinations of both of these approaches (Malaisé et al., 2006; Yi, 2008; Nikolova et al., 2010). There is also an important distinction to be made between the data content of a taxonomy and the methods used for visualisation. There have been user studies which have focussed on evaluation different visualisations while keeping the data constant (Katifori et al., 2007). In this paper we are considering only the data content, the concepts and relations, so visualisations are kept constant.

User studies are certainly important. However these studies often don't answer certain questions about the taxonomy. We can measure aspects of their experience but we might not have a fine-grained understanding of which aspects of the taxonomy were positively or negatively perceived. Also such studies require users to be physically present at a machine specially set up for logging. This can make such studies expensive and time-consuming. A different approach is to use intrinsic evaluations. Yu et al. (2007) present a wide range of ontology evaluation approaches as applied to variations of the Wikipedia category structure such as

[4] http://www.nlm.nih.gov/mesh/

fanout, tangledness, relationship richness, class richness, importance connectivity and cohesion. Cohesion attempts to measure the degree to which similar items are clustered together under a single node in the taxonomy. This was initially proposed by (Boyd-Graber et al., 2009) and has also recently been applied to cultural heritage (Hall et al., 2012).

Another evaluation is to measure the accuracy of the child-parent pairs from the taxonomies by asking evaluators to classify them as either *isa* or *notisa* relations (Snow et al., 2004; Ponzetto and Strube, 2011). A similar method is used here but expanded to give a more detailed understanding of the types of the relations found in the different taxonomies.

3 Resources and tools

This section lists some key resources and tools that are used in this paper.

3.1 Wikipedia Miner

The Wikipedia Miner (Milne and Witten, 2008) tool is used in this paper both as a tool to help map items into existing taxonomies and as a way to generate a novel taxonomy from scratch. Wikipedia Miner is a 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. Disambiguation of terms within the text is performed first. A machine-learning classifier is used with several features. The main features used are *commonness* and *relatedness*, as in Medelyan et al. (2008). The commonness of a target sense is defined by the number of times it is used a destination from some anchor text e.g. the anchor text 'Tree' links to the article about the plant more often than the mathematical concept and is thus more common. Relatedness gives a measure of the similarity of two articles by comparing their incoming and outgoing links. The performance achieved using their approach is currently state of the art for this task. The Wikipedia Miner software is freely available[5].

3.2 Europeana cultural heritage collection

Cultural heritage items from Europeana are used for the evaluation. Europeana is a large online aggregation of cultural heritage collections from across Europe. In this paper a snapshot of the English subset of the data from March 2011 is used. This comprises 547780 items in total. Each item consists of an XML metadata record. This comprises a number of fields the most informative of which are `dc:title`, `dc:subject`, `dc:description` which contain the title, subject keywords and a textual description of the item. About 74% of the items have an associated image which is displayed on the portal website. A difficulty with this collection is that a significant number of the items have very little associated metadata. In the worst case some items have only a one-word title, with no subject or description. This problem is dealt with implicitly in the methods described below, where such sparse records are effectively filtered out in the mapping and taxonomy generation stages and are not included in the evaluations.

4 Taxonomies and mappings

Six taxonomies were tested in these experiments. Four of these were based on existing taxonomies which have been mostly manually created: the Library of Congress Subject Headings, WordNet Domains, Wikipedia Taxonomy and DBpedia. The remaining two taxonomies were automatically derived from the metadata present for the items in the collections: WikiFreq and

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

LDA topics. This section gives a description of each of these taxonomies and how the items in the collection were mapped into them. Statistics for each taxonomy are presented at the end of this section.

4.1 Manually created taxonomies

4.1.1 Library of Congress Subject Headings (LCSH)

The LCSH comprises a controlled vocabulary maintained by the United States Library of Congress for use in bibliographic records. They are used in many libraries to organise their collections as well as for organising materials online.

The text from the `dc:subject` field in the Europeana item are used for the mapping. The text is lemmatized using Freeling (Padró, 2011). 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.

4.1.2 WordNet domains

The WordNet hierarchy is a fine-grained classification which is too detailed for browsing, with more than a hundred thousand nodes and concepts like *entity, natural phenomenon* and *body of water*. Instead *WordNet domains* organises the WordNet concepts into a smaller hierarchy, with only 164 domain labels which are easily understood by a general user. The domain labels have been semi-automatically applied to each of the synsets in WordNet. Each synset is annotated with each one label from a set of about two hundred. The information provided by the domain labels is complementary to the data existing already 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 polysemy.

WordNet domains lists the domain labels for all open class words in WordNet, but it only contains a few proper nouns. Given the large concentration of proper nouns in Europeana, we extend the list of words using Yago2. 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 again to use the `dc:subject` field to link Europeana items to Yago2 entities (using lemmatization and finding the longest possible match). These are then mapped to the WordNet Domain labels via the Yago2 entity-to-synset and the MCR synset-to-WordNetDomain mappings.

4.1.3 Wikipedia Taxonomy

Wikipedia Taxonomy (Ponzetto and Strube, 2011) is a taxonomy derived from Wikipedia categories. The authors create the Wikipedia Taxonomy by keeping the isa relations between Wikipedia categories and discarding the rest. We first apply Wikipedia Miner (see Section 3.1) over the Europeana items to find the relevant Wikipedia entities in the the `dc:subject` field. Then, we link the Europeana item to all Wikipedia Taxonomy categories which are related to these entities.

4.1.4 DBpedia ontology

The DBpedia ontology (Auer et al., 2007) is a small, shallow ontology manually created based on information derived from Wikipedia. Contrary to the previous vocabularies described above, the DBpedia ontology is a formalised ontology, including inference capabilities. The authors provide the instances of each ontology class, i.e. the set of Wikipedia entities pertaining to this class. For mapping Europeana items to DBpedia ontology classes, we first apply Wikipedia Miner to find the relevant Wikipedia entities to the item, and then link the item to the classes these entities belong.

4.2 Automatically created data-driven taxonomies

4.2.1 LDA topic modelling

Latent Dirichlet Allocation (LDA) is a state-of-the-art topic modelling algorithm, that creates a mapping between a set of topics T and a set of items I, where each item $i \in I$ is linked to one or more topics $t \in T$. Each item is input into LDA as a bag-of-words and then represented as a probabilistic mixture of topics. The LDA model consists of a multinomial distribution of items over topics where each topic is itself a multinomial distribution over words. The item-topic and topic-word distributions are learned simultaneously using collapsed Gibbs sampling based on the item - word distributions observed in the source collection (Griffiths and Steyvers, 2004). LDA has been used to successfully improve result quality in Information Retrieval (Azzopardi et al., 2004; Wei and Croft, 2006) tasks and is thus well suited to support exploration in digital libraries.

To turn the flat LDA topic model into a navigable hierarchy, Griffiths and Tenenbaum (2004) describe a hierarchical LDA approach. However this was found to be prohibitively time consuming given our large data-set. Instead a recursive divide and conquer approach was used, which was much more efficient. The number of topic groups at each stage was limited to a maximum of 9 to make the hierarchy manageable for users to navigate. The algorithm is outlined below.

1) Run LDA over the corpus to determine the document-topic probabilities. The number of topics $topic\_n$ to generate is automatically determined using this equation:

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

2) For each topic used the document-topic probabilities that LDA outputs to identify the set of documents associated with that topic. Each document is assigned only to its highest-probability topic. While this removes some of the power inherent in the LDA topic model, we believe that from a navigational perspective it is better if each document is located at only one point in the hierarchy and not at multiple points. To give each topic a label, we simply selected the highest-probability word from each topic's topic-word distribution.

3) If a topic set has less than 60 items then stop. Otherwise go back to 1) using the set of items identified in 2) as the corpus.

4.2.2 Wikipedia link frequencies

This is a novel method for taxonomy creation which uses Wikipedia article links as the concept nodes in the taxonomy. The first step is to add inline article links to all the item texts in the collection using Wikipedia Miner (see Section 3.1). A confidence threshold of 0.5 was used to help ensure the links were of high quality - that is they are correctly disambiguated and relevant to the topic of the item.

The first step is to find the frequency counts of all article links that occur in the items. Let L be the set of all links found in the items. Then the frequency function $F : L \rightarrow \mathbb{N}$ gives the global frequency count for occurrences of the link in all items.

The following procedure is then used for each item to create and populate the taxonomy. Let $S \subset L$ be the set of links found in that item. The links are ordered in S by order of frequency according to the F function (most frequently occuring first) to give an ordered list of links $a_1, a_2, a_3 \cdots a_n$. The item is then inserted into the tree under the branch $a_1 \rightarrow a_2 \rightarrow a_3 \cdots \rightarrow a_n$, with a_1 at the top level in the tree and the item appearing under the node a_n. If this branch does not already exist in the tree then it is created.

It was found that using this approach the branching factor was very high at some levels so the number of child nodes at each level was limited to at most 20. Furthermore only items with at least 2 links were used to prevent a large number of single-linked items appearing at the top level. Concepts with less than 20 items were also filtered out.

This method is labelled WikiFreq in the remainder of the paper.

4.3 Taxonomy statistics

Table 1 shows some statistics for each taxonomy:

- The number of items that are mapped into the taxonomy.
- The average number of parents for each item.
- The average depth from the root node to an item.
- The number of top level nodes in the taxonomy.

A problem with some of the manual taxonomies is the very high number of top level nodes, which makes it difficult for users to browse. However there is no obvious way to select suitable top level nodes in these taxonomies. Additionally some of the taxonomies assign items to many parent nodes - this means that the data is repeated across the taxonomy. This is not a problem in itself, but is likely to mean that items may often be assigned to incorrect nodes.

5 Experiments

Two evaluations were performed on the taxonomies. The first measured the cohesion of the item clustering, and the second gathered human judgements of the relations that were found between child-parent concept pairs in the taxonomy. For both evaluations online surveys were created using an in-house crowdsourcing interface. Links to the surveys were sent out to a mailing list comprising thousands of students and staff members at the University of Sheffield.

Type	Taxonomy	Items	Nodes	Avg. parents	Avg. Depth	Top nodes
Manual	LCSH	99259	285238	1.8	1.97	28901
	DBpedia	178312	273	4.2	2	30
	Wiki Taxonomy	275359	121359	11.7	1.13	10417
	WN domains	308687	170	7.1	7.1	6
Automatic	LDA topics	545896	22494	1	7.3	9
	Wiki Freq	66558	502	1	3.39	24

Table 1: Statistics for each taxonomy

5.1 Cohesion

A cohesive cluster is defined as one in which the items are similar while at the same time clearly distinguishable from items in other clusters (Tan et al., 2006). To measure the cohesiveness of the taxonomies we use the *intruder detection* task originally devised in Boyd-Graber et al. (2009) and recently used for cultural heritage items in Hall et al. (2012). The idea of this is to present 5 items to an evaluator. Four of these are taken from one concept node in the taxonomy and the other (the *intruder*) is randomly picked from elsewhere in the taxonomy. The more cohesive the concept in the taxonomy the more obvious it should be which is the intruder item. Each *unit* was displayed as a list of five images along with the titles of the items. An example[6] of a cohesive unit is shown in Figure 1. To generate good quality units for the evaluation the *informativeness* of items was calculated as follows:

$$\begin{aligned} informativeness(item) = \quad & length(item_{title})/avgL_{title} * log(N/count(item_{title})) \\ & + length(item_{desc})/avgL_{desc} * log(N/count(item_{desc})) \\ & + length(item_{subj})/avgL_{subj} * log(N/count(item_{subj})) \end{aligned}$$

where $title, desc, subj$ refer to the title, description and subject fields of the metadata, $avgL_X$ is the average length of field X over the whole collection, $length(item_X)$ is the length of that item field text, and $count(item_X)$ gives the frequency of that item field text over the whole collection. The higher the resulting value the more informative the item is. Note that as well as taking into account the length of the fields, this also weights by the inverse document frequency (idf) value, so very frequently occuring terms will be downweighted.

The most informative items were selected for each category. This helped to ensure that the users were presented with informative items, allowing them to have enough information to decide which was most likely to be the intruder. The same also applies to the taxonomy mappings being evaluated; it is difficult to correctly map items with very little information in the metadata. The procedure for selecting the sample units was as follows:

[6] Images reproduced from Wikipedia and subject to relevant licenses.
http://en.wikipedia.org/wiki/File:York_Minster_close.jpg.
http://en.wikipedia.org/wiki/File:Wells_Cathedral,_Wells,_Somerset.jpg.
http://en.wikipedia.org/wiki/File:West_Side_of_Westminster_Abbey,_London_-_geograph.org.uk_-_1406999.jpg.
http://en.wikipedia.org/wiki/File:Goatfell_from_Brodick_Harbour.jpg
http://en.wikipedia.org/wiki/File:Sfec-durham-cathedral-2007-263.JPG

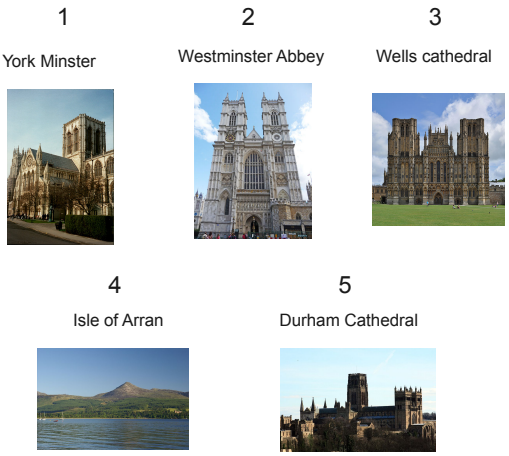

Figure 1: Example of a cohesive unit. Here the intruder item is number 4.

1. Select categories at random that have at least 4 items.

2. For each category:

   (a) Return the 4 items in the category which were most informative.

   (b) To find the intruder item, select 100 items at random from the whole collection and return the most informative item.

Six units were shown on each page, one of which was always a control unit. The control units were manually chosen to be examples where the intruder was very obvious. The purpose of the control units was to ensure the quality of the judgements, since if a participant got the control unit wrong it was an indication that they were not taking the task seriously.

Thirty non-control units were created from each taxonomy. Altogether 134 people attempted the survey. 23 of the users evaluated at least one control unit wrong, or evaluated less than 5 units in total, and so their answers were excluded. The remaining 111 participants contributed a total of 1255 answers. Each unit received a minimum of 5 answers and an average of 6.97 answers. A unit was judged as cohesive if more than 80% of the annotators agreed on the same intruder.

Type	Taxonomy	Coherent units	Percentage
Manual	LCSH	19	63.3
	DBpedia	17	56.7
	Wiki Taxonomy	18	60.0
	WN domains	15	50.0
Automatic	LDA topics	17	56.7
	Wiki Freq	**29**	**96.7**

Table 2: Number of coherent units (out of 30) for each of the taxonomies.

The results (Table 2) show that most of the taxonomies achieved roughly the same level of cohesion for the clusters, roughly between 50 and 63%. However the WikiFreq taxonomy performed far better, with only one unit of the 30 judged as not coherent. This success shows that the Wikipedia links are very effective as a means of grouping together similar items. This might be explained by considering that items grouped together under the same node will share a number of keywords which link to the same Wikipedia articles which would ensure that the items are very similar. In contrast the Wikipedia taxonomy and DBpedia ontology use categories rather than articles in Wikipedia as the concept nodes. These are much more loosely defined; each article in Wikipedia can belong to many categories and each category contains many articles. The results also indicate that Wikipedia articles as entities are much more clearly defined than the LDA topic keywords and thus work much better at grouping together the similar items.

5.2 Relation classification

Previous work has evaluated taxonomies by presenting child-parent pairs of concept nodes to evaluators and asking them a simple boolean question - does the pair represents a valid hypernymic relation, i.e. is it true that "ChildNode isa ParentNode"? (Ponzetto and Strube, 2011; Snow et al., 2004). We would expect the manually created taxonomies to perform well here. The automatic methods also intend to create a hierarchical structure, with more general concepts at the top nodes going to more specific in the lower nodes.

Here we conduct a slightly deeper analysis of what kinds of relations were present in these taxonomies. Instead of a simple boolean question a two-part question was used. Given a child-parent pair A, B the evaluators were asked two questions:

1. Are the two concepts A and B related? (Yes/No/I don't know) The evaluators were asked to judge the relation within the context of the cultural heritage taxonomy. A positive example was presented: Westminster and London, which were related because Westminster is in London. A negative example was Fish and Bicycle which were unrelated and would not be a useful pair to include in a taxonomy.

2. If Yes, then how would you best define the relationship? Is A more specific than B, less specific than B, neither, or don't know? Examples were also given to help with this question. Westminster is *more* specific than London since Westminster is within London. The term Scientist is less specific than Physicist, since while all Physicists are Scientists, not all Scientists are Physicists (they could be biologists or chemists for example). For

Type	Taxonomy	Child (A)	Parent (B)
Manual	LCSH	Work	Human Behaviour
		Braid	Weaving
		Time	Geodetic Astronomy
	DBpedia	Mountain Range	Place
		Fern	Plant
		Congressman	Politician
	Wiki Taxonomy	Mammals of Africa	Wildlife of Africa
		Schools in Wiltshire	Schools in England
		British Culture	European Culture
	WN domains	vehicles	transport
		mechanics	engineering
		home	applied science
Automatic	LDA topics	earthenware	dish
		view	church
		tunnel	chapel
	Wiki Freq	Corrosion	Coin
		Interior Design	Industrial Design
		Towpath	Waterscape

Table 3: Some examples of child-parent pairs from each taxonomy.

the 'neither' option consider Physicist and Biologist. The concepts are related (both are scientists) but neither is more specific than the other.

Forty non-control pairs from each taxonomy were presented to the evaluators giving a total of 240 pairs. Examples of concept pairs from each taxonomy are shown in Table 3. As for the previous experiment control pairs were manually identified where the answer should be obvious. Five pairs were shown on each page of which one was always a control pair.

Altogether 270 people attempted this survey. 97 people evaluated more than half the control pairs wrong or evaluated less than 5 pairs in total, and so their answers were excluded. Of the 173 remaining participants, a total of 3826 evaluations were made for each pair. A minimum of 8 evaluations and an average of 15.94 evaluations were made for each instance.

Type	Taxonomy	Yes	No	Don't know	Agreement
Manual	LCSH	74.2	8.8	17.0	79.1
	DBpedia	86.6	11.2	2.2	88.4
	Wiki Taxonomy	**96.1**	1.7	2.3	**95.9**
	WN domains	77.1	14.5	8.4	83.9
Automatic	LDA topics	30.3	**50.3**	19.3	71.6
	Wiki Freq	47.6	16.5	**35.8**	70.9

Table 4: Are *A* and *B* related?

The results for the relatedness question (Table 4) show a clear pattern. The manually created taxonomies are markedly more likely to contain clearly related pairs of concepts. The Wikipedia

taxonomy hierarchy is the highest performing in this regard which suggests that the user-created category hierarchy is of high quality and has easily understood concepts and relations. DBpedia scores slightly lower. This difference might be explained due to DBpedia containing article entities as well as categories. The lower score suggests either that articles are not always placed in the best categories, or that it is harder for users to identify article-category relationships. WordNet domains scores lower. This may be due to the domain concepts being sometimes quite general and possibly harder for general users to understand (for example one pair of concepts was 'color' and 'factotum'). LCSH scored surprisingly low considering that it is a manually created taxonomy. This suggests that the concepts and relations in this hierarchy are even harder for users to understand or identify. Finally the two data derived taxonomies score lower still. For the WikiFreq taxonomy a high percentage of the relations were classified as 'Don't know'. This may be because a high number of the article links are about quite obscure concepts which most people would not know about. Finally the LDA topics produced the highest number of definite 'No' judgements which shows that the taxonomy may be difficult or confusing for users to navigate.

Type	Taxonomy	$A < B$	$A > B$	Neither	Don't know	Agreement
Manual	LCSH	65.4	8.7	23.4	2.5	68.7
	DBpedia	76.2	4.9	18.1	0.7	78.9
	Wiki Taxonomy	**78.3**	4.7	16.0	0.9	**82.8**
	WN domains	63.6	6.3	28.0	2.0	67.6
Automatic	LDA topics	21.4	14.8	**62.1**	1.6	61.0
	Wiki Freq	30.9	**22.6**	43.6	**2.9**	67.0

Table 5: Specificity of the pairs, with A the child node, and B the parent node. $A < B$ means A is more specific than B.

The results for the specificity question are shown in Table 5. These follow a roughly similar pattern to the relatedness. The $A < B$ case is the most desirable for the taxonomies since we would prefer the most general concepts at the top of the hierarchy narrowing down into more specific concepts. The Wikipedia taxonomy and DBpedia both score relatively highly here, although both contain a surprisingly high number of cases where neither A or B was identified as more specific than the other (16.0 and 18.1% respectively). For the Wikipedia taxonomy this shows that although almost all child-parent pairs are considered to be related concepts, they are not always easily identified with the child as more specific than the parent. Both WordNet domains and LCSH fare worse, again with more relations identified as 'neither'. The WikiFreq taxonomy contains a more mixed set of results with quite a high proportion of relations the 'wrong way round' with A deemed to be less specific than B, although the highest number falls into the 'neither' category. This result is a reflection of the nature of the links within the items. The taxonomy is ordered with the most frequent occuring links at the top going down towards the least. Clearly this is not enough to create the kind of general-to-specific relationships which are desirable. Finally the results for the LDA topics show that the majority are defined as 'neither' - the concepts are topically related but mostly without any specificity ordering.

Conclusion and perspectives

Developing effective taxonomies for the purpose of organising large number of data items is a complex task. Existing manually created taxonomies might be accurate and well structured but may not be adequate for a specified domain or may be hard for users to navigate. Automatic methods for deriving taxonomies have the advantage of closely reflecting the nuances of the data - but organising the derived concepts into meaningful relations remains a problem.

The experiments in this paper shows some surprising results. The LCSH taxonomy has been manually created for the purpose of organising library collections and so might be the obvious choice to organise CH data online. However the results show that the relations within LCSH are defined less clearly than that of the Wikipedia derived taxonomies. WordNet domains performs at a similar level to the LCSH in terms of the quality of the relations. The LDA topic hierarchy gave poor results in terms of the identified topics. The topic pairs were often unrelated, and had no general-to-specific structure as would be desirable for this application.

The WikiFreq hierarchy performed slightly better in this regard. Just over half the concept pairs were judged to be related. Just under a third were labelled as 'Don't know' which may reflect the obscurity of the concept nodes identified. It was hoped that organising the frequency counts of the links would organise the hierarchy into a general-to-specific direction. This was not achieved, although the hierarchy does have the benefit of providing the user with an overview of the collection by immediately seeing which kind of items are most prevalent.

In terms of cohesion all the taxonomies achieved similar results except for the WikiFreq taxonomy which achieved almost perfect cohesion. This shows how effective the Wikipedia links are in grouping together similar items.

It is also worth noting that WikiFreq and LCSH map significantly fewer items into the taxonomy.

Future work will continue with different evaluation approaches, such as domain/task coverage and accuracy of the mappings. We also aim to expand the evaluations to include user studies; a key question is how well these taxonomies assist users when used for browsing large collections, such as Europeana. The aim is to see if there is a correlation between the intrinsic results that were found here with the extrinsic quality judgements when used in real life applications. A promising line of work will be to build on the WikiFreq approach by integrating with the high quality Wikipedia taxonomy knowledge base. The hope is that using this approach will generate highly coherent units along with a well structured conceptual tree.

Acknowledgements

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. This research was also partially funded by the Ministry of Economy under grant TIN2009-14715-C04-01 (KNOW2 project).

References

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

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

Azzopardi, L., Girolami, M., and Van Rijsbergen, C. (2004). Topic based language models for ad hoc information retrieval. In *Neural Networks, 2004. Proceedings. 2004 IEEE International Joint Conference on*, volume 4, pages 3281–3286. IEEE.

Bentivogli, L., Forner, P., Magnini, B., and Pianta, E. (2004). Revising the wordnet domains hierarchy: semantics, coverage and balancing. In *Proceedings of the Workshop on Multilingual Linguistic Ressources*, pages 101–108. Association for Computational Linguistics.

Boyd-Graber, J., Chang, J., Gerrish, S., Wang, C., and Blei, D. (2009). Reading tea leaves: How humans interpret topic models. In *Proceedings of the 23rd Annual Conference on Neural Information Processing Systems*.

Fellbaum, C., editor (1998). *WordNet: An electronic lexical database*. MIT Press.

Griffiths, D. and Tenenbaum, M. (2004). Hierarchical topic models and the nested chinese restaurant process. In *Advances in neural information processing systems 16: proceedings of the 2003 conference*, volume 16, page 17. The MIT Press.

Griffiths, T. and Steyvers, M. (2004). Finding scientific topics. *Proceedings of the National Academy of Sciences of the United States of America*, 101(Suppl 1):5228.

Hall, M. M., Clough, P. D., and Stevenson, M. (2012). Evaluating the use of clustering for automatically organising digital library collections. In *Theory and Practice of Digital Libraries 2012*.

Hearst, M. (1992). Automatic acquisition of hyponyms from large text corpora. In *Proceedings of the 14th conference on Computational linguistics-Volume 2*, pages 539–545. Association for Computational Linguistics.

Hearst, M. (2009). *Search user interfaces*. Cambridge Univ Pr.

Hoffart, J., Suchanek, F., Berberich, K., Lewis-Kelham, E., De Melo, G., and 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*, pages 229–232. ACM.

Katifori, A., Halatsis, C., Lepouras, G., Vassilakis, C., and Giannopoulou, E. (2007). Ontology visualization methods - a survey. *ACM Comput. Surv.*, 39(4).

Malaisé, V., Aroyo, L., Brugman, H., Gazendam, L., de Jong, A., Negru, C., and Schreiber, G. (2006). Evaluating a thesaurus browser for an audio-visual archive. In Staab, S. and Svátek, V., editors, *Managing Knowledge in a World of Networks*, volume 4248 of *Lecture Notes in Computer Science*, pages 272–286. Springer Berlin / Heidelberg. 10.1007/11891451_25.

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

Medelyan, O., Witten, I. H., and Milne, D. (2008). Topic Indexing with Wikipedia. In *Proceedings of the Association for the Advancement of Artificial Intelligence (AAAI) WikiAI workshop*.

Milne, D. and Witten, I. H. (2008). Learning to Link with Wikipedia. In *Proceeding of the 17th ACM conference on Information and Knowledge Management*, pages 509–518.

Nikolova, S., Ma, X., Tremaine, M., and Cook, P. (2010). Vocabulary navigation made easier. In *Proceedings of the 15th international conference on Intelligent user interfaces*, IUI '10, pages 361–364, New York, NY, USA. ACM.

Padró, L. (2011). Analizadores multilingües en freeling. *Linguamatica*, 3(2):13–20.

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

Sanderson, M. and 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*, pages 206–213. ACM.

Snow, R., Jurafsky, D., and Ng, A. (2004). Learning syntactic patterns for automatic hypernym discovery. *Advances in Neural Information Processing Systems 17*.

Tan, P., Steinbach, M., Kumar, V., et al. (2006). *Introduction to data mining*. Pearson Addison Wesley Boston.

Wei, X. and 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*, pages 178–185. ACM.

Wilson, M., B, K., MC, S., and B, S. (2010). From keyword search to exploration: Designing future search interfaces for the web. *Foundations and Trends in Web Science*, 2(1):1–97.

Yi, M. (2008). Information organization and retrieval using a topic maps-based ontology: Results of a task-based evaluation. *Journal of the American Society for Information Science and Technology*, 59(12):1898–1911.

Yu, J., Thom, J., and Tam, A. (2007). 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.