Using Ontological and Document Similarity to Estimate Museum Exhibit Relatedness

KARL GRIESER and TIMOTHY BALDWIN, The University of Melbourne
FABIAN BOHNERT, Monash University
LIZ SONENBERG, The University of Melbourne

Exhibits within cultural heritage collections such as museums and art galleries are arranged by experts with intimate knowledge of the domain, but there may exist connections between individual exhibits that are not evident in this representation. For example, the visitors to such a space may have their own opinions on how exhibits relate to one another. In this article, we explore the possibility of estimating the perceived relatedness of exhibits by museum visitors through a variety of ontological and document similarity-based methods. Specifically, we combine the Wikipedia category hierarchy with lexical similarity measures, and evaluate the correlation with the relatedness judgements of visitors. We compare our measure with simple document similarity calculations, based on either Wikipedia documents or Web pages taken from the Web site for the museum of interest. We also investigate the hypothesis that physical distance in the museum space is a direct representation of the conceptual distance between exhibits. We demonstrate that ontological similarity measures are highly effective at capturing perceived relatedness and that the proposed RACO (Related Article Conceptual Overlap) method is able to achieve results closest to relatedness judgements provided by human annotators compared to existing state-of-the art measures of semantic relatedness.

Categories and Subject Descriptors: I.7.1 [**Document and Text Processing**]: Document and Text Editing—*Document management*

General Terms: Experimentation, Measurement, Human Factors

Additional Key Words and Phrases: Ontological similarity, document similarity, museum exhibit, Wikipedia, WordNet

ACM Reference Format:

Grieser, K., Baldwin, T., Bohnert, F., and Sonenberg, L. 2011. Using ontological and document similarity to estimate museum exhibit relatedness. ACM J. Comput. Cult. Herit. 3, 3, Article 10 (March 2011), 20 pages.
DOI = 10.1145/1921614.1921617 http://doi.acm.org/10.1145/1921614.1921617

1. INTRODUCTION

The research presented in this article identifies the different ways in which museum visitors categorize relationships between exhibits. We use this both as the basis of the development of an exhibit ontology, and in the categorization of individual exhibits in a given museum space. As a first step towards this goal, we focus on the numeric estimation of exhibit relatedness via a range of methods, and use a set of exhibit relatedness judgements from museum visitors to evaluate each. The primary means of

Authors' addresses: K. Grieser; email: kgrieser@csse.unimelb.edu.au; T. Baldwin; email: tb@ldwin.net; F. Bohnert; email: fabian.bohnert@monash.edu; and L. Sonenberg; email: l.sonenberg@unimelb.edu.au.

DOI 10.1145/1921614.1921617 http://doi.acm.org/10.1145/1921614.1921617

exhibit relatedness estimation we consider are: (1) document similarity, based on either Wikipedia or museum Web site documents; (2) taxonomic similarity, based on Wikipedia document categories; and (3) physical distance in the museum space. In doing this, we draw on research from a diverse range of sources, including psycholinguistics, lexical similarity, ontologies, text mining, and geospatial indexing. In particular, we examine the effect of collaboratively constructed ontologies with regard to conceptual relatedness of museum exhibits as opposed to professionally curated ontologies used in previous studies.

Recently there has been a move towards providing visitors to museums and other cultural heritage (CH) spaces with personalized tours [Fink and Kobsa 2002; Schmidt-Belz et al. 2002]. These tours can be created explicitly by a visitor prior to entering the collection, or tailored to a visitor while browsing a collection. In order to create a dynamic tour for a given visitor, there is the need to: (1) model a visitor's preferences [Bohnert et al. 2009; Fink and Kobsa 2002; Zukerman and Albrecht 2001], and (2) have knowledge about the content of individual exhibits and connections between pairs of exhibits [Aroyo et al. 2007; Cox et al. 1999; Grieser et al. 2007]. The focus of this article is on the second of these requirements: using Web documents to represent museum exhibits, and ontological and document similarity to model the strength of relationships between exhibits.

Museum galleries are generally designed around a common theme (e.g.,*Living and Dying* or *How Things Fly*), and professionally curated so that exhibits are organized in a coherent fashion relevant to that theme, with closely related objects in close physical proximity to each other. For example, artifacts from the same era or of the same function are often presented together. The task of tour personalization can be seen as one of matching the interests of a visitor to the themes represented in the museum. However, visitors to a museum or other CH site can categorize the museum space in a way particular to the context of their visit (e.g., preferring to visit more tactile exhibits to entertain small children, or choosing to visit all exhibits from a particular location or era, irrespective of theme). That is, they often have their own opinions on the degree of relatedness of exhibits, independent of the themed design of a gallery or exhibition.

Traditional lexical ontologies provide a framework from which to analyze the structure of a lexicon. For alternate ontologies, the methods for computing semantic relatedness and semantic similarity can be adapted as long as the hierarchical structure of the ontology allows this [Ponzetto and Strübe 2007a]. This approach can achieve comparable performance over standardized datasets, as demonstrated in Ponzetto and Strübe [2007b]. However, while computational measures have been developed to identify the relationships between words or documents, there has been no analysis of the manner in which people regard the relationships between real-world objects.

The primary contributions of this research are as follows.

—We successfully use Wikipedia document categories as a conceptual graph, and apply lexical similarity methods to them.

—We propose an encyclopedic similarity measure, taking into account the category membership of documents and their interlinkedness.

—We demonstrate the feasibility of exhibit similarity modelling in a situated context over a novel dataset.

—We investigate the relationship between the curators' and museum visitors' views of exhibit relatedness, and demonstrate significant differences.

—We show that the proposed RACO (Related Article Conceptual Overlap) measure of encyclopedic similarity outperforms current state-of-the-art measures of ontological and document similarity for a data set of exhibit descriptions.

While various computational methods have been developed to identify relationships between documents or words (e.g., Rubenstein and Goodenough [1965] and Ponzetto and Strübe [2007a]), there is currently no standard method of identifying the manner in which people view the relationships among real-world objects. This research demonstrates a novel approach, using taxonomic and document-based methods of semantic relatedness methods to estimate museum exhibit relatedness.

2. BACKGROUND

2.1 Exhibit Associations in Cultural Heritage Collections

Cultural Heritage spaces such as historical sites and museums are providing greater access to their collections through portable assistants [Benelli et al. 1999; Oppermann and Specht 2000], wearable computing [Randell and Muller 2002; Vlahakis et al. 2002], digital augmentation of museums [Brogni et al. 1999], and virtual museums and showcases [Aroyo et al. 2007; Cox et al. 1999; Rowe and Barnicoat 2009]. This has enabled museums to reach wider audiences and to better communicate the importance of their collections. Digital access to these collections has also enabled the development of route planning applications, as well as exploration of a collection prior to or after a visit. These tours of the museum can be selected from a number of premade paths that focus on a central theme of the exhibits along the path, much as a human tour guide would (e.g., the Louvre's Thematic Trails[1]), or they can be created to center around a visitor's personal interests (e.g., the Kubadji project at Melbourne Museum, Australia and the CHIP project at the Rijksmuseum, The Netherlands). The use of a central theme that connects the exhibits to one another provides the visitor with a deeper understanding of the overall importance of a collection, and why they are placed in the same gallery or exhibition.

If a tour is rigidly defined and does not allow for alterations, then a visitor may not be able to continue with the tour if they are interested in something outside the scope of the tour. Similarly, if a tour has been created around a visitor's interests, it can be difficult to define a central theme, or connect the exhibits together into a coherent narrative or tour, ultimately leaving the visitor without a clear impression of any central theme (other than their own interests) that the exhibits share. It may even be the case that a visitor has chosen to focus on a single gallery in a museum because that is where they believe their interests lie, and while this gallery may be interesting to the visitor, they remain unaware of the other exhibits in the museum that also fit their interests. For example, a visitor to Melbourne Museum interested in fossils and dinosaurs would be most interested in seeing the Evolution Exhibition in the Life and Science Gallery. However, there are many exhibit areas that directly relate to the theme of dinosaurs and Earth's geologic history, such as the series of exhibits on the breakup of Gondwana present in the Forest Gallery. Cox et al. [1999] demonstrated a solution to this problem by building on a model of visitor interests as the tour progressed. In this case the domain was a digital museum. By identifying what attributes of an exhibit a visitor had asked for more information on, future exhibits viewed by the same visitor generated natural language text that described content in relation to the previously interesting items.

The personalization of content through digital collections has been a major focus of many CH projects [Aroyo et al. 2007; Benelli et al. 1999; Cox et al. 1999]. Personalization allows a user to feel as though the information they receive is directly applicable to their interests, even if the contribution of the user is as superficial as defining the aesthetic aspects of the information presentation. Personalization can be performed by explicitly asking the user what sort of information they are interested in, or how they want information to be presented to them. An alternate approach is to identify what the user is interested in based on how they act, and then using that behavior to create a model of that

[1] http://www.louvre.fr/llv/activite/liste_parcours.jsp?bmLocale=en

user's interests. Examples of this implicit user modelling include the Microsoft Windows Start Menu, which contains a list of the user's recently used applications, amazon.com's product recommendations, which present products that the user may be interested in purchasing based on previous items they have viewed [Linden et al. 2003], and the models of visitors' interest in exhibits developed within the Kubadji project [Bohnert and Zukerman 2009]. These models can be constructed by using the behavior of other users (collaborative filtering), or when the behavior of users is too dissimilar or idiosyncratic, to identify connections within the information itself (content modelling) [Zukerman and Albrecht 2001]. Much personalization is done by recommender systems [Resnick and Varian 1997], which use both collaborative filtering [Hill et al. 1995] and content modelling [Cox et al. 1999], or even a combination of the two [Basu et al. 1998], to identify additional sources or pieces of information that may be of interest to the user. This article presents a content-based method of identifying related exhibits.

There is a trend towards personalizing the information presented on museum Web sites to individual museum visitors [Bowen and Filippini-Fantoni 2004]. In many cases, the information presented to the visitor pertains directly to the visitor's physical interaction with the museum by either aiding the visitor in planning a future museum tour [Wang et al. 2008] or reviewing a previous visit [Mulholland et al. 2005]. While the delivery of personalized content to museum visitors as they visit a museum has been questioned because of the distraction it causes from the physical environment of the museum [Filippini-Fantoni and Bowen 2007], numerous studies demonstrate the benefit of using personalized content to aid in memory retention [Cox et al. 1999], education [Anderson et al. 2000; Roussou 2004] and user engagement [Woodruff et al. 2002]. Marty [2007] argues that the process of continually reviewing previous tours and planning for future tours aids a visitor's understanding of the collection and relationships within it. Rather than distracting a visitor from the visiting experience, it allows the visitor to more fully enjoy the space as it is presented. Distraction may be caused by the process of continually pushing information onto the visitor when they may not be in search of it [Cheverst and Smith 2001]. When the visitor is given the option of selecting when to access the extra information, it removes this distraction factor.

Specific examples of content personalization in identifying relationships between exhibits that are interesting to the visitor (both the exhibits themselves, and the relationships) have used the attributes of previously rated exhibits to identify other exhibits that the visitor might find interesting (e.g., Aroyo et al. [2007] and Cox et al. [1999]). Grieser et al. [2007], on the other hand, used common attributes of exhibits in the current visit to predict future exhibits the visitor might visit, while Bohnert et al. [2008, 2009] used the amount of time a visitor spent viewing exhibits to infer a visitor's interests and pathways. The content-based models explored by Grieser et al. and the collaborative models proposed by Bohnert et al. have the advantage of being nonintrusive, as they do not require explicit exhibit ratings. Additionally they do not rely on the structure of the exhibits when identifying connections between them, overcoming the restriction of requiring exhibits having similar structure in order to compare their qualities.

The identification of reasons for visitors finding commonality between exhibits is a key step in personalizing a tour. Previous studies have used common attributes to align exhibits and identify similarities (e.g., same artist or same style of jewelery, as described by Cox et al. [1999]). This has lead to the use of ontological frameworks as a basis for these comparisons (e.g., The Getty AAT, Iconclass, or the CIDOC Conceptual Reference Model). This methodology is based on the human mind's use of ontological relationships to help organize data and relationships. An example of an ontological representation of human thought is the mind map, which can be used to facilitate learning and memory by identifying how concepts relate to each other. In CH sites such as art galleries, where all exhibits have the same attributes, this method can be appropriate. However, for CH sites that have exhibits of differing backgrounds (e.g., natural history museums or national parks) this method does not adequately account for the diversity of the exhibit structure. Another drawback of using an expert-curated ontology from which to extract relationships is that the structure is defined by a select group of highly informed users, and the relationships created may not be relationships that the layperson understands. Additionally, visitors may not make direct correlations between attributive qualities of exhibits, and may be more interested in features such as color, exhibit content, or the relationship of exhibits with an external topic (e.g., paintings related to the French Revolution). We address this gap by using an alternative semistructured data source that is able to identify relationships between concepts within its hierarchy, through semantic relationships and information content. These semantic relationships facilitate the creation of content-based models for museum collections. In addition to recommending museum exhibits to museum visitors as they tour the museum, the extracted semantic relationships can facilitate the creation of post-visit summaries or activities (e.g., in the case of a school excursion).

Estes [2003] showed that for concepts that do not have a common conceptual frame (or physical structure), people relate concepts using a process of integration. An integrative relationship is the interaction that occurs between two concepts. This is different from attributive comparison, where similarity is determined by the qualities that two concepts have in common. This indicates that for CH sites with diverse collections, highly structured data sources and ontologies are insufficient in capturing user-inferred relations between exhibits. Their key failure is that they do not capture the thought process that the average museum visitor goes through (often an integrative relationship), but rather focus on the organizational hierarchy designed by the collection's curators. This is not to say that the curators' placement of exhibits is an incorrect one, only that there may be multiple interpretations of the exhibit space, or multiple relationships between exhibits.

For the purpose of our research, we define an exhibit as a single artifact regarded in isolation from the rest of the collection in which it appears. For example, exhibits at Melbourne Museum are separated into multiple levels of granularity, the lowest of which is a single exhibit that consists of a single object such as a single taxidermized horse, a mummy in its sarcophagus, or a model of a quartz-reef gold mine.

2.2 Semantic Association

As a first step towards understanding how museum visitors integrate exhibits, we focus on exhibit relatedness: the estimation of the degree to which museum visitors feel that a given pairing of museum exhibits is related. In this, we follow research on lexical similarity, drawing heavily on the early psycholinguistic work of Rubenstein and Goodenough [1965]. We use the methodology of Rubenstein and Goodenough to derive a real-world dataset of exhibit relatedness over pairs of exhibits at Melbourne Museum . This provides the gold-standard dataset for evaluation of different methods for estimating exhibit relatedness in this research.

2.2.1 *WordNet.* WordNet [Fellbaum 1998] is a lexical ontology containing several thousand word senses. At its lowest level, WordNet is a connected set of synsets. A synset is a collection of individual word senses that possess synonymous meaning (e.g., the words *forest*, *woodland*, *timberland*, and *timber* comprise a synset in WordNet 3.0). Relationships connecting synsets include meronymy (part-of) and holonymy (composed-of). The hypernym (is-a) relationship forms a word hierarchy with specific terms at its leaves, and general terms towards its root (e.g., *dog* is-a *canine*, *canine* is-a *mammal*). Methods used to identify the relatedness of words are described in Section 4.

WordNet has proven to be a powerful tool for analyzing and interpreting text. It has been used in applications as diverse as summarization (e.g., Bellare et al. [2004]), information retrieval (e.g., Mandala et al. [1998]) and text categorization (e.g., Rosso et al. [2004]). Of particular interest is the Word Sense Disambiguation (WSD) subtask of semantic relatedness: the use of a lexical ontology (WordNet) to

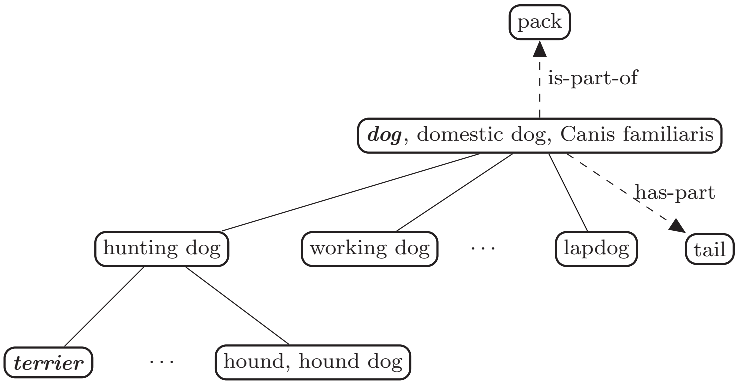

Fig. 1. A section of the WordNet ontology: solid lines between nodes represent is-a (hypernym/hyponym) relationships, and dashed lines represent is/has-part (meronym/holonym) relationships. Each node in the graph represents a single WordNet **synset**; note that a synset can contain a single word, or multiple terms.

determine the semantic relatedness of words within a given context. WSD is the task of determining the sense of an ambiguous word in text. For example, the word *dog* can be interpreted as a noun (the animal) or a verb ("to harass or follow").[2] In the sentence *"A terrier is a kind of dog,"* it is evident that we are referring to the animal. However it is unclear to a computer program which sense we are referring to. By noting that the word *terrier* also occurs in the sentence, one can use the fact that the first (and only) sense of *terrier* is a hyponym of the noun sense of *dog*. The verb sense of *dog* may also possess a relationship with the first sense of *terrier*. However, it is related to a much lesser degree than the noun sense. Computational measures of semantic relatedness and similarity are designed to automate this process of word sense disambiguation. An example of the WordNet structure relating to this example appears in Figure 1.

2.2.2 *Lexical Similarity*. The lexical similarity dataset of Rubenstein and Goodenough (and other datasets of a similar style) has provided the basis for evaluation of a wide range of lexical similarity methods, commonly over lexical ontologies [Budinatsky and Hirst 2005; Ponzetto and Strübe 2007b]. Most such methods have been developed specifically over WordNet, but are applicable to any ontology that provides hierarchical ISA-style links. The use of a gold standard dataset against which all newly developed methods are compared is essential to a field that undergoes constant change. It allows for the definition of benchmarks and gives a purpose to the development of a field by providing a tangible example of why a method is useful. In the case of the Rubenstein and Goodenough dataset, the objective was to determine the similarity of words. The participants of the experiment were given pairs of words ranging from nearly synonymous (e.g., *gem* and *jewel*, or *forest* and *woodland*) to completely unrelated (e.g., *rooster* and *voyage*, or *automobile* and *wizard*). The participants were then asked to rate how similar in meaning these two words were on an integer scale of 0 to 4. The participants were chosen from a general student population. Hence, while they were familiar with the English language they were not necessarily experts in its use. For this study we use a similar set of participants: people who are familiar with a museum, but not necessarily experts on its content. We ask the visitors to rate the relatedness between exhibits rather than the similarity of words.

[2] There are also multiple senses of the noun *dog* including "an unattractive woman," and "a tool fitting." But for this example we only consider the basic noun and verb senses of *dog*.

The concept of semantic relatedness can be adapted to the cultural heritage domain for the purpose of identifying the relatedness of museum exhibits. In previous examples (see Section 2.1) all connections between exhibits are presented to a user. This is valid, but when recommending possibly related exhibits to a user, presenting all candidates no matter how vague the relationship can result in user fatigue. We are able to use the information associated with museum exhibits to measure their relative relatedness and filter out low quality candidates. This can be done by using an ontology of museum exhibits to describe the conceptual relationships between museum exhibits and exploiting this structure to simulate the conceptual relationships that museum visitors think of when regarding museum exhibits, rather than the process of simple attributive comparison suited for use in collections of exhibits with similar structure (described in Section 2.1). Ontologies are organized structures of information consisting of individual entities connected by defined relationships. Ontologies can be used to represent lexical entities and relationships (e.g., WordNet), encyclopedic knowledge (e.g., Cyc), or the organization of living things (e.g., the Linnaean Taxonomy). Ontologies can be planned, with a defined root node and clear hierarchy, or can be unplanned networks of nodes that evolve as needed with new entities and relationships added ad hoc.

2.2.3 *Wikipedia and Semantic Relatedness.* In this research, we will use a nonexpert data source that provides relationships between highly different entities, and that is able to represent the information at a common visitor level: Wikipedia. In recent years, Wikipedia has been increasingly used in document processing tasks, due to its sheer size, multilinguality and domain diversity. Extensive conceptual similarity experiments have been performed in other studies, such as the ones discussed in Gabrilovich and Markovitch [2007] and Milne et al. [2007]. Particularly interesting is the category hierarchy, as each article must be a member of at least one category. This hierarchy has been investigated by Ponzetto and Strübe [2007a, 2007c] as a parallel to other existing hierarchies such as WordNet. In Wikipedia, the articles are created with the intention of being understandable to all users, and even its place in the category hierarchy is reached through discussion and consensus. As a result, an article's content and its organization are designed to make sense to the majority of people viewing it.

In many cases, the alignment methodology used to equate entities within an external resource with Wikipedia concepts is highly dependent on the external resource itself. As an encyclopedic ontology, Wikipedia possesses many articles that have the same name, but different qualities. For this purpose, disambiguation articles exist to list all the possible meanings of an article title. Bunescu and Paşca [2006] used these disambiguation articles to resolve ambiguous named entities. The approach taken by Sarjant et al. [2009] when using Wikipedia to augment Cyc (a relational knowledge database) was to align multiple Wikipedia articles (based on article synonyms and redirects) to Cyc concepts based on the context in which they appeared. Attributive qualities and link structure of Wikipedia articles can also be used to classify the organization of existing ontologies (e.g., Reiter et al. [2008]).

We use Wikipedia as the source of both an ontology of CH artifacts (via the document categories), and encyclopaedic knowledge about individual artifacts. Wikipedia is particularly well suited to this research for a number of reasons. First, it is collaboratively authored and thus presents an educated nonexpert's view of concepts. Hence, we expect it to be a good proxy of the visitor's view of museum exhibits, rather than the prescriptive curators' view. Second, it has considerably broader conceptual coverage than WordNet, especially over named entities (which make up a significant portion of museum exhibits, e.g., TUTANKHAMUN DEATH MASK or ROSETTA STONE). Third, while WordNet contains short lexicographic glosses of each synset, individual Wikipedia articles contain detailed encyclopaedic information on a given topic in a form which is more comparable with the style of presentation in a museum context. For example, the Wikipedia article on FOREST contains detailed information on the distribution of forests, as well as a classification of forests, and discussion of forest management. WordNet, on the other hand, categorizes forest into two noun synsets, glossed briefly as “the trees and other plants in a large densely wooded area” and “land that is covered with trees and shrubs,” Fourth, we can make use of the Wikipedia document categories to achieve an ontological view of the conceptual content of Wikipedia. Comparing WordNet relations with Wikipedia document categories, the first sense of the noun FOREST, for example, has hypernyms including VEGETATION and meronyms including UNDERBRUSH, whereas in Wikipedia, FOREST, FOREST WORKING GROUP, and OLYMPIC NATIONAL FOREST are all members of the category FORESTS.

Using an ontology derived from the Wikipedia document categories, we demonstrate that a metric based on overlapping category membership of articles related to the exhibit outperforms state-of-the-art semantic similarity measures [Jiang and Conrath 1997; Leacock et al. 1998; Lin 1998; Palmer and Wu 1995] over the same dataset. Additionally, we contrast this ontological view of Wikipedia with a document similarity method based on the content of the Wikipedia article associated with a given exhibit. We also use the physical distance between exhibits as an estimate of their conceptual distance, based on the observation that museum galleries are generally designed to flow conceptually, with closely related objects physically located near one another. We calculate the physical distance in two ways: (1) relative to their exact positioning in the museum space, via their shortest-path *exhibit-to-exhibit* walking distance within the physical museum space; and (2) relative to the gallery they are located in, via the shortest-path *gallery-to-gallery* walking distance in the physical museum space.

3. HUMAN JUDGEMENTS OF EXHIBIT RELATEDNESS

In order to generate gold-standard exhibit relatedness data from the perspective of a museum visitor, we distributed a questionnaire to members of Melbourne Museum in the form of an online survey. The survey presented each participant with a series of pairs of exhibits from the Melbourne Museum collection. Each exhibit was a single artifact considered independently from the collection in which it appeared, and presented as a single artifact to the survey participant without describing where it appeared in the museum. Each participant was asked to grade the relatedness of the exhibit pair on an integer scale of 0 to 4, whether they had encountered the two exhibits before, and to optionally provide the reason for the exhibits being related. Forty exhibits from Melbourne Museum 's collection were used in this survey, and each survey participant was asked to compare 15 randomized pairs.

In excess of 500 responses to the questionnaire were received over a one month period in late 2007, each respondent analyzing an average of 15 pairs (the participant had the option to finish early, or to compare more exhibit pairs). There were 7198 exhibit pairs compared, 48% with textual justifications for the relatedness score supplied. In the research described here, we only make use of the numeric relatedness scores. We intend to use the textual justifications in future research on the nature of exhibit relatedness. In the 7198 exhibit pair responses, there were 198 malicious comparisons[3] that greatly increased the variance of some relatedness scores. These comparisons were excluded from the results.

The survey included three fixed exhibit pairings to provide an indication of the level of consensus between participants. The mean and standard deviation of these pairs can be seen in Table I. The standard deviations in these pairings are high and indicate that the participants of the survey did not share a common opinion of what exhibits were related. However, a casual observation of the exhibit names demonstrates the highly different content present in these exhibits, from living things (the ANT COLONY exhibit) to mechanical devices (the RIFLES and PISTOLS exhibit). It may be that

[3] We define a malicious user as one that rates all exhibit pairs they are shown with the same score (i.e., a user whose scores have zero variance).

Table I. Average Relatedness Scores (Mean) and Standard Deviations (s.d.) of Exhibit Pairs Given to all Survey Participants

Exhibit Pair	Relatedness	
	Mean	s.d.
Gorilla Diorama – Trilobite Fossil	1.53	1.18
Rifles & Pistols – Sauropod Bone	0.50	0.88
Ant Colony – Gold Mine Model	1.89	1.32

some participants were able to identify connections between these items based on personal experience, while others were not. By contrast, for example, the outcome of a pairing such as HUMAN SKELETON MODEL and HUMAN ANATOMY MODEL (both present in this survey) is very different, producing a mean score of 3.90, with a standard deviation of 0.30.

3.1 Interannotator Agreement

It should be noted that due to the high variance of relatedness scores from the visitor survey, there is an added difficulty in predicting scores. If visitors are unlikely to agree with each other, then creating a generalized measure to identify the degree of relatedness as viewed by the average visitor is a futile exercise. We define the interannotator agreement as the average of the Pearson correlation between the set of scores of each survey respondent and the average of the other respondents' scores. The resulting interannotator agreement between participants is $\rho = +0.507$. This demonstrates that, even with high variance for some pairwise relatedness scores, there is agreement between the visitors. This also indicates the level of success that can be expected from RACO (Related Article Conceptual Overlap) and other similarity measures (see Section 4), suggesting that correlations of $\rho = 0.4$ can be viewed as a success, given the initial variance.

4. ESTIMATING EXHIBIT RELATEDNESS

We propose three main approaches to estimating the relatedness of a given pairing of museum exhibits: (1) ontological similarity based on Wikipedia document categories; (2) document similarity based on the textual similarity between the Wikipedia articles associated with each exhibit; and (3) physical distance in the museum space. In this section, we detail the individual methods employed for each approach.

4.1 Ontological Similarity

We use the document categories in Wikipedia to generate an ontology of Wikipedia articles, and apply a representative set of lexical similarity measures over the derived graph.

Many of the ontological similarity methods we use assume that the ontology has a unique root, and that a given pair of articles has a unique lowest common subsumer (LCS). The unique root is necessary to ensure that it is possible to reach a unique LCS for each exhibit pair. The LCS is important because it defines the point in the ontology at which the two exhibits are able to define an aspect that they have in common. In some cases the LCS will be very deep in the ontology, indicating that the two exhibits have a highly specific relationship. In others the LCS may be closer to the root of the tree meaning that they can only be related on very broad or unspecific terms. If the two exhibits share nothing in common, their LCS will be the root node of the ontology. Without a unique root node, some exhibit pairs may not have an LCS. As the root, we use the FUNDAMENTAL category. However, Wikipedia does not guarantee a unique LCS. Due to the branching subsumption of categories, there may be multiple categories that two exhibits share as part of their ancestry. Hence, we require a workaround.

Our approach is to use the Floyd-Warshall algorithm to precompute the shortest path between each pairing of Wikipedia articles, and in the case of multiple LCS candidates, select the category of greatest depth.[4] The Floyd-Warshall algorithm [Floyd 1962] determines the shortest path between all pairs of nodes in a directed graph (such as the Wikipedia Category hierarchy) by determining the distance from a node to its parent to be 1, and then utilizing this distance information to find all paths of length equal to 2, and so forth until the distance between all pairs has been found. Because of the large size of the Wikipedia Category Hierarchy, this algorithm is desirable because of the relatively low complexity ($O(n^3)$, where n is the number of nodes in the graph, or in our case, the number of documents in Wikipedia).

In keeping with standard practice in the lexical similarity literature [Budinatsky and Hirst 2005; Ponzetto and Strübe 2007b], we experiment with two basic approaches to ontological similarity calculation in addition to our proposed RACO measure (Section 4.2): path-based (Section 4.3) and information-based (Section 4.4) ontological similarity.

4.2 Related Article Conceptual Overlap

Analysis of the survey responses indicates that the majority (58%) of responses describing the reasons for two exhibits being related explained the element that was common to both exhibits (e.g., "Both have trees," "Both animals"). A smaller subset (35%) identified an integrative reason for the exhibits being related (e.g., "A place to live" as the relationship between the TARANTULA and BURNT TREE exhibits).

These two relationships can be identified in Wikipedia: the aspect that is common to both exhibits is the category that both articles fall under, while the integrative relationships are the links to other articles within an article's body text (called *outlinks*). By combining integrative and attributive similarity, we define a measure that is able to outperform state-of-the-art measures of ontological similarity.

We use Wikipedia's link structure and category membership to identify the strength of the relationship between articles by examining the category overlap of related articles. While this method is inspired by the survey responses, it makes no use of the survey results. We are therefore able to compare the performance of the proposed method against the gold-standard survey response data without fear of bias.

The number of categories that are common to the sets of categories of outlinked articles of two articles, a and b, can be defined is as follows:

$$\text{Category-Overlap}(a, b) = \left| \left(\bigcup_{p \in O(a)} C(p) \right) \bigcap \left(\bigcup_{p \in O(b)} C(p) \right) \right|, \tag{1}$$

where $O(a)$ is the set of outlinks from article a and $C(p)$ is the set of categories of which article p is a member.

As long articles have more outlinks than shorter articles (and hence a greater number of article supercategories) it is necessary to normalize the resulting score by the total number of outlink article supercategories to reduce the bias towards giving larger articles greater scores due to their size. We use Dice's coefficient [Dice 1945] to normalize the similarity measure. The final form of the Related Article Conceptual Overlap (RACO) method is as follows:

$$sim_{RACO}(a, b) = \frac{2 \times \left| \left(\bigcup_{p \in O(a)} C(p)\right) \bigcap \left(\bigcup_{p \in O(b)} C(p)\right) \right|}{\left|\bigcup_{p \in O(a)} C(p)\right| + \left|\bigcup_{p \in O(b)} C(p)\right|}. \tag{2}$$

[4] Similar to the approach adopted for WordNet, where there is similarly no guarantee of a unique LCS, due to multiple inheritance.

Table II. Overall Pearson Correlation (ρ) and Statistical Significance (p-value) Between the Gold Standard Relatedness Scores and the Various Exhibit Relatedness Estimation Methods (the Highest Correlation is Bold-Faced)

	Overall
Ontological Similarity	
RACO	+**0.404** (1.6×10^{-33})
Shortest-Path	+0.212 (7.5×10^{-10})
Shortest-LCS-Path	+0.133 (1.3×10^{-4})
Leacock-Chodorow	+0.263 (1.9×10^{-14})
Wu-Palmer	+0.009 (7.9×10^{-1})
Lin	+0.007 (8.4×10^{-1})
Jiang-Conrath	−0.022 (5.3×10^{-1})
Document Similarity	
tf·idf (Wikipedia)	+0.209 (1.6×10^{-9})
tf·idf (Museum)	+0.294 (6.6×10^{-8})
Physical Distance	
Exhibit distance	+0.196 (1.5×10^{-8})
Gallery distance	+0.144 (3.2×10^{-5})
Upper Bound	+0.507

Table III. Pearson Correlation (ρ) and Statistical Significance (p-value) Between the Gold Standard Relatedness Scores and the Various Exhibit Relatedness Estimation Methods (for Exhibit Pairings of Differing Physical Distance, Based on 3-Class Equal-Frequency Discretization; the Highest Correlation is Bold-Faced in Each Column)

	Near	Mid-distance	Far
Ontological Similarity			
RACO	+**0.448** (5.8×10^{-15})	+**0.304** (3.0×10^{-7})	+**0.379** (9.0×10^{-11})
Shortest-Path	+0.230 (1.2×10^{-4})	+0.162 (7.4×10^{-3})	+0.223 (2.0×10^{-4})
Shortest-LCS-Path	+0.085 (1.6×10^{-1})	+0.154 (1.1×10^{-2})	+0.097 (1.1×10^{-1})
Leacock-Chodorow	+0.285 (2.2×10^{-6})	+0.212 (4.1×10^{-4})	+0.253 (2.3×10^{-5})
Wu-Palmer	−0.033 (5.9×10^{-1})	+0.107 (7.7×10^{-2})	+0.015 (7.9×10^{-1})
Lin	−0.007 (9.1×10^{-1})	+0.096 (1.2×10^{-1})	−0.057 (3.6×10^{-1})
Jiang-Conrath	−0.031 (6.2×10^{-1})	−0.090 (1.5×10^{-1})	+0.057 (3.5×10^{-1})
Document Similarity			
tf·idf (Wikipedia)	+0.225 (1.8×10^{-4})	+0.169 (5.0×10^{-3})	+0.155 (1.0×10^{-2})
tf·idf (Museum)	+0.318 (7.3×10^{-4})	+0.053 (5.9×10^{-1})	+0.060 (5.4×10^{-1})
Physical Distance			
Exhibit distance	+0.262 (1.1×10^{-5})	+0.069 (2.6×10^{-1})	+0.040 (5.1×10^{-1})
Gallery distance	+0.162 (7.0×10^{-3})	+0.066 (2.7×10^{-1})	+0.084 (1.7×10^{-1})
Upper Bound	+0.531	+0.414	+0.537

This form of the equation is the one used to produce the results in Table II and Table III.

4.3 Path-Based Ontological Similarity

Path-based ontological similarity measures consider all edges in the Wikipedia category hierarchy to have unit weight, and simply count the number of edges in the path between a given pairing of articles. We explore the following variants.

Shortest-Path. The simplest approach to path-based ontological similarity is simple shortest path (Shortest-Path): the count of the number of nodes in between two given nodes.[5]

Shortest-LCS-Path. The (SHORTEST-LCS-PATH) method is an extension of Shortest-Path , where we first identify the LCS for a given pair of articles, and return the distance between the articles via the LCS. Note that the shortest path through the LCS can be different from the simple shortest path because of our method for disambiguating the LCS via the distance to the root.

Leacock-Chodorow. The Leacock-Chodorow measure of similarity [Leacock et al. 1998] scales the shortest path between two nodes ($sp(a, b)$) by the maximum depth of the hierarchy (D): the maximum edge count to the root across all Wikipedia articles:

$$sim_{lch}(a, b) = -\log \frac{sp(a, b)}{2 \cdot D}. \tag{3}$$

Wu-Palmer. The Wu-Palmer measure [Palmer and Wu 1995] scales the depth of each node (l_a and l_b, respectively) to their LCS ($lcs_{a,b}$) by the depth of the LCS:

$$sim_{wup}(a, b) = \frac{2 \cdot depth(lcs_{a,b})}{l_a + l_b + 2 \cdot depth(lcs_{a,b})}. \tag{4}$$

The depth of a node is defined as the shortest path distance from the node to the root node via the category hierarchy.

4.4 Information Content-Based Ontological Similarity

Information content-based ontological similarity measures weight edges in the ontology by an estimate of the relative semantic difference between the concepts they represent. This is conventionally interpreted by the synset priors, based on analysis of the token frequency of senses, for example, relative to a text corpus such as SemCor. In our case, we instead used the results of an independent tracking experiment to observe the prior for a nonmember Melbourne Museum visitor to visit a given exhibit. The assumption is that exhibits appearing in the same visitor tour are alike because the visitor is interested in them both in some way, similar to the assumption used in word sense disambiguation that words that appear in the same sentence share information. The information content of a term is calculated by taking the inverse log of the probability of it occurring in a corpus:

$$IC(t) = -\log p(t). \tag{5}$$

Using an information theoretic interpretation of the exhibit-visit priors, we estimate exhibit similarity based on the following two methods.

Lin. The Lin measure [Lin 1998] scales the Information Content (IC) of the two nodes by the information content of their LCS, indicating the amount of information the two nodes share.

$$sim_{lin}(a, b) = \frac{2 \cdot IC(lcs_{a,b})}{IC(a) + IC(b)}. \tag{6}$$

[5] Strictly speaking, this is a distance metric not a similarity measure, but we label it as an ontological similarity method for terminological convenience (similarly for Shortest-LCS-Path , and for the measures of physical distance presented in Section 4.6). When calculating the Pearson correlation between a distance-based method and the gold standard, we reverse the sign of the resultant ρ value. This is due to the comparison of a similarity score with a dissimilarity (distance) score, and is done for ease of comparison with other similarity-similarity correlations.

Jiang-Conrath. The Jiang-Conrath measure [Jiang and Conrath 1997] utilizes the information content of the two nodes and their LCS to determine their similarity.

$$sim_{jcn}(a,b) = \frac{1}{IC(a) + IC(b) - 2 \cdot IC(lcs_{a,b})}. \tag{7}$$

4.5 Document Similarity

Document similarity was calculated based on the cosine similarity of the documents associated with each of the two exhibits, interpreting the article as a term vector with *tf·idf* weighting [Salton and McGill 1983]. Here, we use two document sources: (1) Wikipedia documents (denoted *tf·idf* (Wikipedia)); and (2) the exhibit descriptions present on the Melbourne Museum Web site (denoted *tf·idf* (Museum)).

The frequency of a term (t_i) for a document (d_j) is defined as the number of times the term occurs in the document, divided by the sum of number of occurrences of all terms t_k in the document d_j:

$$tf_{i,j} = \frac{n_{i,j}}{\sum_k n_{k,j}}. \tag{8}$$

The inverse document frequency is defined to be the logarithm of the quotient of the total number of documents ($|D|$) divided by the number of documents containing the term:

$$idf_i = \log \frac{|D|}{|\{d_j : t_i \in d_j\}|}. \tag{9}$$

The *tf·idf* term weight is the product of these two values: $tf{\cdot}idf_{i,j} = tf_{i,j} \times idf_i$.

The simple cosine similarity of the document vectors is insufficient as it does not take into account the frequency of terms across a large document set. This means that documents that contain very different content could be scored as highly similar because they have many words in common that are not representative of their content (e.g., words such as *the*, *and*, *or*, etc.). By normalizing the frequency of a term within a document by its frequency across the entire document set (by using *tf·idf* weighting), the impact of very frequent and common terms is reduced.

4.6 Physical Distance

The museum space is carefully themed by curatorial staff, placing semantically related exhibits in spatially cohesive groups, for example, in galleries and exhibitions. Hence physical distance between exhibits and galleries can be viewed as the museum curators' viewpoint of the collection. However, as curators are domain experts and highly familiar with their collection, a visitor may have a different opinion of how the exhibits relate to one another. For example, museum galleries and exhibitions often bring together exhibits with a common theme such as a Dinosaur gallery or a Flight exhibition. In these cases, exhibits that are highly related are also physically close. Thus, we use physical distance as a measure of relatedness to provide a baseline for the preceding measures of similarity.

In this article, we use two measures of distance derived from the physical layout: one derived from the actual locations of the exhibits (*exhibit distance*), and one derived from the arrangement of galleries (*gallery distance*). To calculate the distances, we employed an SVG-file based representation of the museum, mapped onto a graph structure which preserves the physical layout of the museum (preventing paths from passing through walls or ceilings). Exhibit distances are calculated on this graph using the physical location of the exhibits in the museum space. To obtain gallery distances, we first identified the physical centroids of the galleries. We then determined the gallery-to-gallery distance for each exhibit pair as the distance between the centroids of the galleries the exhibits are placed in (in the case of two exhibits being in the same gallery, the distance between them was set to zero).

5. EXPERIMENTS

In order to evaluate the different methods, we calculated the Pearson correlation (as used in similar studies, e.g., Ponzetto and Strübe [2007b]) between each of our methods (Section 4) and the set of gold standard relatedness scores (Section 3).

In addition to evaluating the overall performance of a given method, we calculated its performance over three equal-frequency bands of physical exhibit-to-exhibit distance. This was in order to investigate the hypothesis that there is an inverse relationship between physical distance and relatedness (the further apart two exhibits are, the less related they become), as intended by the curators' layout of the exhibits into galleries and exhibition areas. The 820 exhibit pairs were partitioned equally into the three bands: 274 pairs in the near band and 273 pairs in both the mid-distance and far bands.

For the purposes of calculating document similarity, we manually aligned each exhibit used in the survey (described in Section 3) with the Wikipedia article that best described its content. Of the 40 exhibits, 26 had an equivalent Web page at the Melbourne Museum Web site; these documents were also used to compute the similarity of exhibits. Results presented for document similarity over the museum documents are based on this subset of the exhibits. As this is an incomplete dataset, it is less than ideal to use, however it is included to demonstrate the impact of the use of a different data source. When calculating the correlation between gold standard scores and the document similarity using Melbourne Museum Web site documents, only exhibit pairs comprised of the 26 aligned exhibits were considered (resulting in only 325 exhibit pairs). The bandwise separation of this subset of 26 exhibits results in 109 pairs for the near distance band, 108 pairs for the mid-distance band, and 108 pairs for the far band.

5.1 Results

In Table II, we present the Pearson correlation and the two-tailed p-value between the gold standard dataset and the relatedness estimates from each of the proposed methods. We use the physical distance between exhibits as our baseline, as it is an established set of relationships (exhibits placed in the same collection by curators share a common theme). Five measures outperform this baseline (RACO, Shortest-Path, Leacock-Chodorow, and the Document Similarity using Wikipedia articles and Melbourne Museum Web site pages.). However, the RACO measure is nearest to the interannotator agreement upper bound (Section 3.1). The effectiveness of these methods is further examined in the following.

To test the influence that the layout of Melbourne Museum has on visitor judgements of relatedness, we partitioned the scores into three equal bands (near, middistance and far) based on physical distance between exhibits. These split scores were then compared to the gold standard visitor scores which were split in the same fashion (by physical distance between exhibits). The similarity measures may be able to identify when exhibits have a lot in common (i.e., in the same gallery), but for exhibits that are separated by large distances they may be unable to identify any possible relationships. For example, exhibits at opposite ends of the museum are less likely to be related to one another than if they were next to each other. This can be seen in the organization of exhibits and galleries over multiple levels of granularity: exhibits with a common theme are placed together into an exhibition (e.g., a collection of exhibits about the human body), and exhibitions with a common theme are placed in the same gallery (e.g., placing the *Human Body*, *Evolution* and *Insects* exhibitions into the *Science and Life* gallery). The bandwise correlations (Table III) identify for which measures this is the case.

The best correlation achieved for the bandwise comparison is again that of the RACO measure. The RACO measure outperforms all other measures across all three bands, and comes closest to the inter-visitor correlation upper bound for all bands.

6. DISCUSSION

The interannotator agreement derived from the visitor survey (see Section 3) indicates that the connections that individual museum visitors identify between museum exhibits can be diverse. However, even with this diversity of opinion, computational measures of semantic relatedness such as RACO demonstrate a correlation over a sample set of museum exhibits. This section analyzes the results of the experiments performed in Section 5.

6.1 Path-Based Measures

The measures that utilize the simple path distance between nodes (Leacock-Chodorow and Shortest-Path) outperform methods that take into account the deepest common ancestry of the two nodes (Wu-Palmer and Shortest-LCS-Path). This is true for the overall correlation as well as at each level of the bandwise correlations.

Detailed analysis of each article's set of ancestor categories (the set of all subsuming categories) indicates that even for relatively shallow articles, the ancestor set can be as large as 1500 Wikipedia categories. When using the shortest path to the root node to define a category's depth, there are many instances where a category is subsumed by a deeper category (due to the parent category having a longer path to the root node). Additionally, the branching structure of the category hierarchy means that there is a large subset of nodes that are common to all articles' ancestor sets. The combination of these two factors (categories that are deeper than their children, and a large set of categories common to all articles' ancestor sets) creates a situation where the LCSs arrived at for all exhibit article pairs is a set of approximately 5 categories. In the majority of cases, this node is the same: the category SYNAPSIDS. The next two most commonly occurring LCSs are CARL JUNG and TERRITORIES BY LANGUAGE, with only two pairs reaching LCSs of the desired specificity and relatedness to the original articles (these categories were PASSERI and COELUROSAURS). The removal of these hub nodes produces an identically aligned new set of hub nodes (meaning that there exists a large set of categories at that depth), and the removal of a hub node just means selecting the next in the list of candidate nodes at the desired level. Hence, without further analysis of how the subsuming category relates to the two articles on an informational, rather than solely ontological level, the use of the LCS to identify relatedness across the Wikipedia Category hierarchy is inappropriate.

By placing the restriction that the LCS of two articles can only be shallower than the shallower of the two articles, a new set of LCSs arises. This set of LCSs produces a $\rho = +0.119$ correlation for the basic LCS measure, and a $\rho = +0.004$ for Wu-Palmer. Analysis of the LCSs shows that there is a single LCS for all pairs that meet at a level. The impact of defining the maximum depth for the LCS of two articles simply produces a hub category for each level. For example, the LCS of a pair of articles of depths 8 and 5 will always be at depth 4 (one node shallower than the shallowest of the pair) no matter what the pair; furthermore it will always be the same category no matter what the starting articles. Without specifically examining the information that is associated with the category and the articles, using the LCS to compute article association will be unsuccessful, as it will always be the same node for a given depth.

6.2 Information Content-Based Measures

The performance of the information content-based measures has the potential to exceed the performance of the path-based and document-similarity measures, as they directly take into account the probability that exhibits occur in the same visit. This insight is only gained with the use of an appropriate corpus (in this case, a set of paths through Melbourne Museum). It is evident that the corpus of visitor paths did not provide the correct insight into common groupings of exhibits. The criteria of first-time visitors to the museum meant that the visitors were unfamiliar with the exhibit space and considered all locations equal, resulting in a distribution that may have given a higher information content to more visible exhibits and a lower information content to more obscure exhibits.

The category subsumption problem present with the LCS-based measures is again a cause of the failure of these methods. With the majority of categories being present in all articles' ancestor nodes, the IC of many categories approaches 1. Thus the IC of many LCSs provides no additional information in calculating either the Jiang-Conrath or Lin scores.

6.3 Document Similarity

In contrast to the information content-based measures, the measures of document similarity demonstrate that a correlation exists between article content and opinions of exhibit relatedness. Both document similarity measures outperform the physical similarity baseline over all pairs, and the *tf·idf* (Museum) measure shows the second highest correlation for the near band. This is an unsurprising result when one considers that the documents have been tailored to the specific context of the museum (and hence contain some information that the visitor may have already encountered in the museum space). This performance is not present in either the mid-distance or far bands, where the *tf·idf* scores using Wikipedia documents shows a higher correlation with the survey pairs. The cause of this anomaly is that documents for exhibits in a given gallery (which inevitably end up in the near band of exhibit pairings) are generally authored by the same gallery curators, and hence are a better reflection of the relative exhibit relatedness. As we cross gallery boundaries however, this effect disappears, and the general-purpose Wikipedia documents are a superior representation of exhibit content. Additionally, the relative sparsity of documents present on the Melbourne Museum Web site makes it less preferable to Wikipedia when being used for further exhibit alignment.

The two measures of document similarity exclusively examine the content of the documents, not taking into account the organization of these documents. Even with this restriction, the *tf·idf* measures outperform the majority of ontological measures. However, by regarding direct links between articles and their ontological grouping, RACO offers a higher correlation with the exhibit relatedness scores.

6.4 Analysis of Bandwise Correlation Behavior

Unsurprisingly, the physical distance based methods performed better for near exhibit pairings (where the exhibits are generally in the same gallery) than for mid-distance and far pairings. This supports our hypothesis that physical distance is an effective representation of exhibit relatedness only within galleries. What was surprising was that the best performing of the ontological similarity and document similarity methods at each band of physical distance outperformed the best of the physical distance-based methods. This is not an indictment of the exhibits' placement, nor of the use of physical distance as a similarity measure. Indeed, when presented in isolation, exhibits are regarded without the influence of surrounding related exhibits, whereas the intention of the curator is to place exhibits into areas encompassing a common theme. There was no hint of an overarching theme presented to the Melbourne Museum members when their judgements were made, and there was thus the potential for exhibits to be judged less related than if they had been encountered in the actual museum space.

Comparing the best of the established state-of-the-art ontological similarity methods (Leacock-Chodorow) with the best of the document similarity methods (*tf·idf* (Museum)), the difference in overall correlation is surprisingly small, given that Leacock-Chodorow does not have direct access to any information specific to the context of the exhibits in Melbourne Museum. Additionally, Leacock-Chodorow is more consistent across the different distance bands, and is applicable to all exhibit pairings (recall that not all exhibits had a Web page on the Melbourne Museum Web site). However, our proposed RACO measure outperforms both of these state-of-the-art measures. Analysis of the bandwise separation of scores (Table III) demonstrates that the RACO measure is superior to all other tested measures across respective bands. The band with the smallest correlation using the RACO measure (the mid-distance band at $\rho = 0.304$) achieves comparable performance to the next best correlation across all bands (the near band of *tf·idf* (Museum) at $\rho = 0.318$).

One can notice a pattern in the bandwise correlation obtained for the path-based measures: measures using the shortest path distance (Shortest-Path and Leacock-Chodorow) do not have as high a correlation with visitor judgements in the mid-distance band as at the near and far distance bands. This can be explained by the measures being able to identify when a pair is highly similar or dissimilar, but mid-distance exhibits having fewer elements in common because the exhibitions containing the two exhibits are both members of a larger gallery (e.g., Dinosaurs and Human Anatomy exhibitions are both members of the Natural Science gallery).

6.5 Portability

Another key feature of our demonstrated implementation (RACO) is that it is founded on Wikipedia and the exhibits' alignment to Wikipedia articles. This allows other exhibits to be aligned to articles, either from the same collection or an entirely different museum or cultural heritage site. The size and versatility of Wikipedia also allows this methodology to be extended to domains other than cultural heritage. An example of the use of aligned online data sources is the Netflix Prize.[6] The Netflix Prize was a machine-learning task that had participants construct recommender systems designed to accurately predict a user's rating of (and hence potential interest in) films. Many entries used online encyclopedias such as the Internet Movie Data Base[7] (IMDB) to identify connections between films. The article relationships present in Wikipedia can be used to identify relationships between films beyond simple genre or director alignment as with an attributive database such as IMDB.

7. CONCLUSION

This article demonstrates that the task of identifying the degree of relatedness between museum exhibits can be approached similarly to the task of lexical similarity, based on the existence of a broad-coverage ontology such as is provided by the Wikipedia category hierarchy. We utilize the link structure and category membership of Wikipedia to measure the strength of relationships between articles. Our evaluation data took the form of relatedness scores between pairs of museum exhibits in an experiment similar to the Rubenstein and Goodenough [1965] word-pair experiment. By aligning a collection of museum exhibits to Wikipedia articles, we measured the correlation between the gold standard relatedness scores (provided by museum visitors) and relatedness scores produced by a set of state-of-the-art similarity measures. The methods of Shortest-Path, Leacock-Chodorow (both using Wikipedia's category hierarchy) and document similarity using Wikipedia articles, exceeds the performance of the baseline metric of physical distance between the exhibits within Melbourne Museum. Importantly, the proposed RACO measure of article similarity over Wikipedia that utilizes the category overlap of article outlinks outperforms these state-of-the-art similarity measures, and approaches the upper-bound correlation determined by the interannotator agreement.

The arrangement of the exhibits at Melbourne Museum is based on thematic areas, placing exhibits with common content together. This placement best represents the classification of exhibits from a curatorial viewpoint, and in many cases these exhibits are constructed specifically to fit within a gallery theme. However, even when created specifically for a gallery or exhibition, exhibits still maintain connections with exhibits in other collections within the museum, or even in other cultural heritage sites. By using measures such as RACO to identify exhibits that are closely related, personalized tours can

[6] http://www.netflixprize.com
[7] http://www.imdb.com

be created which center around visitors' interests. This is a novel approach to the problem of identifying levels of relatedness between museum exhibits in that it builds upon an established taxonomy of encyclopedic knowledge.

Projects such as CHIP [Aroyo et al. 2007] and ILEX [Cox et al. 1999] require knowledge of the common attributes and qualities of exhibits to determine the relatedness. However, when the exhibits being compared lack a common conceptual frame, direct attributive comparison is difficult [Estes 2003].

By extending the ability to identify how much two exhibits have in common, it is possible to identify the specific relationships that relate exhibits to one another. RACO identifies what categories and links two articles share, and it is a simple progression to identify the points the articles have in common. The application of such a method has many uses in cultural heritage recommender systems. For example, if a museum visitor is interested in a specific aspect of an exhibit and wishes to find out more about it, knowledge of how other exhibits relate to this exhibit will allow a recommender system to identify the subset of exhibits that possess the semantic relationship in line with the visitor's interests. This information may be explicitly sought for by the visitor, or presented in a post-visit summary that mentions how the exhibits the visitor encountered relate to other exhibits. Such information is also useful in identifying a central theme or trend in a visitor's tour. The task of identifying the specific reasons for two exhibits being semantically related is a separate task from the task of relatedness strength, but it is a task that the authors intend to explore in future work.

REFERENCES

ANDERSON, D., LUCAS, K. B., AND GINNS, I. S. 2000. Development of knowledge about electricity and magnetism during a visit to a science museum and related post-visit activities. *Sci. Edu. 84*, 658–679.

AROYO, L., BRUSSEE, R., RUTLEDGE, L., GORGELS, P., STASH, N., AND WANG, Y. 2007. Personalized museum experience: The Rijksmuseum use case. In *Online Proceedings of Museums and the Web*.

BASU, C., HIRSH, H., AND COHEN, W. 1998. Recommendations as classification: Using social and content-based information in recommendation. In *Proceedings of the 15th National Conference of Artificial Intelligence*. 714–720.

BELLARE, K., SARMA, A. D., SARMA, A. D., LOIWAL, V., ANDU G RAMAKRISHNAN, V. M., AND BHATTACHARYA, P. 2004. Generic text summarization using WordNet. In *Proceedings of the International Conference on Language Resources and Evaluation*. 691–694.

BENELLI, G., BIANCHI, A., MARTI, P., SENNATI, D., AND NOT, E. 1999. HIPS: Hyper-interaction within physical space. In *Proceedings of the IEEE International Conference on Multimedia Computing and Systems*. 1077–1078.

BOHNERT, F., SCHMIDT, D. F., AND ZUKERMAN, I. 2009. Spatial processes for recommender systems. In *Proceedings of the 21st International Joint Conference on Artificial Intelligence*. 2022–2027.

BOHNERT, F. AND ZUKERMAN, I. 2009. Non-intrusive personalisation of the museum experience. In *Proceedings of the 17th International Conference on User Modeling, Adaptation, and Personalization (UMAP)*. 197–209.

BOHNERT, F., ZUKERMAN, I., BERKOVSKY, S., BALDWIN, T., AND SONENBERG, L. 2008. Using interest and transition models to predict visitor locations in museums. *AI Communi. 21,* 2–3, 195–202.

BOWEN, J. P. AND FILIPPINI-FANTONI, S. 2004. Personalisation and the web from a museum perspective. In *Online Proceedings of Museums and the Web*.

BROGNI, B. A., AVIZZANO, C. A., EVANGELISTA, C., AND BERGAMASCO, M. 1999. Technological approach for cultural heritage: Augmented reality. In *Proceedings of the 8th International IEEE Workshop on Robot and Human Interaction*. 206–212.

BUDINATSKY, A. AND HIRST, G. 2005. Evaluating WordNet-based measures of lexical sematic relatedness. *Comput. Ling. 32,* 1, 13–47.

BUNESCU, R. C. AND PAŞCA, M. 2006. Using encyclopedic knowledge for named entity disambiguation. In *Proceedings of the 11th Conference of the European Chapter of the Association for Computational Linguistics*. 9–16.

CHEVERST, K. AND SMITH, G. 2001. Exploring the notion of information push and pull with respect to user intention and disruption. In *Proceedings of the International Workshop on Distributed and Disappearing User Interfaces in Ubiquitous Computing*. 67–72.

COX, R., O'DONNELL, M., AND OBERLANDER, J. 1999. Dynamic versus static hypermedia in museum education: An evaluation of ILEX, the intelligent labelling explorer. In *Proceedings of the Artificial Intelligence in Education Conference*. 181–188.

DICE, L. R. 1945. Measures of the amount of ecologic association between species. *Ecology 26,* 3, 297–302.

ESTES, Z. 2003. A tale of two similarities: comparison and integration of conceptual combination. *Cog. Sci. 27,* 6, 911–921.

FELLBAUM, C., Ed. 1998. *WordNet: An Electronic Lexical Database*. MIT Press, Cambridge.

FILIPPINI-FANTONI, S. AND BOWEN, J. P. 2007. Bookmarking in museums: Extending the museum experience beyond the visit? In *Online Proceedings of Museums and the Web*.

FINK, J. AND KOBSA, A. 2002. User modeling for personalized city tours. *Artif. Intell. Rev. 18,* 1, 33–74.

FLOYD, R. W. 1962. Algorithm 97. *Comm. ACM 5,* 6, 345.

GABRILOVICH, E. AND MARKOVITCH, S. 2007. Computing semantic relatedness using Wikipedia-based explicit semantic analysis. In *Proceedings of the 20th International Joint Conference on Artificial Intelligence*. 1606–1611.

GRIESER, K., BALDWIN, T., AND BIRD, S. 2007. Dynamic path prediction and recommendation in a museum environment. In *Proceedings of the Workshop on Language for Cultural Hetitage Data*. 49–56.

HILL, W., STEAD, L., ROSENSTEIN, M., AND FURNAS, G. 1995. Recommending and evaluating choices in a virtual community of use. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. 194–201.

JIANG, J. AND CONRATH, D. 1997. Semantic similarity based on corpus statistics and lexical taxonomy. In *Proceedings of the International Conference on Research in Computational Linguistics*. 19–33.

LEACOCK, C., CHODOROW, M., AND MILLER, G. 1998. Using corpus statistics and WordNet relations for sense identification. *Comput. Ling. 24,* 1, 147–65.

LIN, D. 1998. Automatic retrieval and clustering of similar words. In *Proceedings of the 36th Annual Meeting of the ACL and 17th International Conference on Computational Linguistics (COLING/ACL)*. 768–774.

LINDEN, G., SMITH, B., AND YORK, J. 2003. Amazon.com recommendations: Item-to-item collaborative filtering. *Internet Comput. 7,* 1, 76–80.

MANDALA, R., TAKENOBU, T., AND HOZUMI, T. 1998. The use of WordNet in information retrieval. In *Proceedings of the Usage of WordNet in Natural Language Processing Systems, COLING/ACL Workshop*. 31–37.

MARTY, P. F. 2007. Museum websites and museum visitors: Before and after the museum visit. *Museum Manag. Curatorship 22,* 4, 337–360.

MILNE, D., WITTEN, I., AND NICHOLS, D. 2007. A knowledge-based search engine powered by Wikipedia. In *Proceedings of the 16th ACM Conference on Information and Knowledge Management*. 445–454.

MULHOLLAND, P., COLLINS, T., AND ZDRAHAL, Z. 2005. Bletchley Park Text: Using mobile and semantic web technologies to support the post-visit use of online museum resources. *J. Interactive Media Edu.* http://jime.open.ac.uk/2005/24/mulholland-2005-24.html.

OPPERMANN, R. AND SPECHT, M. 2000. A context-sensitive nomadic information system as an exhibition guide. In *Proceedings of the Handheld and Ubiquitous Computing Second International Symposium*. 127–142.

PALMER, M. AND WU, Z. 1995. Verb semantics for English-Chinese translation. *Mach. Translat. 10*, 59–92.

PONZETTO, S. AND STRÜBE, M. 2007a. Deriving a large scale taxonomy from Wikipedia. In *Proceedings of the 22nd AAAI Conference on Artificial Intelligence*. 1440–1445.

PONZETTO, S. AND STRÜBE, M. 2007b. Knowledge derived from Wikipedia for computing semantic relatedness. *J. Artif. Intell. Res. 30,* 1, 181–212.

PONZETTO, S. P. AND STRÜBE, M. 2007c. An API for measuring the relatedness of words in Wikipedia. In *Proceedings of the ACL Demo and Poster Sessions*. 49–52.

RANDELL, C. AND MULLER, H. L. 2002. The eSleeve: a novel wearable computer configuration for the discovery of situated information. In *Proceedings of 22nd International Conference on Distributed Computing Systems Workshops*. 793–796.

REITER, N., HARTUNG, M., AND FRANK, A. 2008. A resource-poor approach for linking ontology classes to Wikipedia articles. In *Proceedings of the Conference on Semantics in Text Process.* 381–387.

RESNICK, P. AND VARIAN, H. R. 1997. Recommender systems. *Comm. ACM 40,* 3, 56–58.

ROSSO, P., FERRETTI, E., JIMÉNEZ, D., AND VIDAL, V. 2004. Text categorization and information retrieval using wordnet senses. In *Proceeding of the 2nd International WordNet Conference*. 299–304.

ROUSSOU, M. 2004. Learning by doing and learning through play: An exploration of interactivity in virtual environments for children. *ACM J. Comput. Entertain. 2,* 1, 1–23.

ROWE, P. AND BARNICOAT, W. 2009. NZMuseums: Showcasing the collections of New Zealand te Papa Tongarewa, New Zealand. In *Online Proceedings of Museums and the Web*.

RUBENSTEIN, H. AND GOODENOUGH, J. 1965. Contextual correlates of synonymy. *Comm. ACM 8,* 10, 627–633.

SALTON, G. AND MCGILL, M. J. 1983. *Introduction to Modern Information Retrieval*. McGraw-Hill, New York.

SARJANT, S., LEGG, C., ROBINSON, M., AND MEDELYAN, O. 2009. “All you can eat” ontology-building: Feeding Wikipedia to Cyc. In *Proceedings of IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology*. 341–348.

SCHMIDT-BELZ, B., NICK, A., POSLAD, S., AND ZIPF, A. 2002. Personalized and location-based mobile tourism services. In *Proceedings of the Workshop on Mobile Tourism Support Systems*. 36–42.

VLAHAKIS, V., KARIGIANNIS, J., TSOUTROS, M., IOANNIDIS, N., AND STRICKER, D. 2002. Personalised augmented reality touring of archaeological sites with wearable and mobile computers. In *Proceedings of the 6th International Conference on Wearable Computers*. 15–22.

WANG, Y., SAMBEEK, R., SCHUURMANS, Y., AROYO, L., STASH, N., RUTLEDGE, L., AND GORGELS, P. 2008. Be your own curator with the CHIP tour wizard. In *Online Proceedings of Museums and the Web*.

WOODRUFF, A., AOKI, P. M., GRINTER, R. E., HURST, A., SZYMANSKI, M. H., AND THORNTON, J. D. 2002. Eavesdropping on electronic guidebooks: Observing learning resources in shared listening environments. In *Online Proceedings of Museums and the Web*.

ZUKERMAN, I. AND ALBRECHT, D. 2001. Predictive statistical models for user modeling. *User Model. User-Adapt. Interact. 11,* 1-2, 5–18.

Received July 2009; revised January 2010, April 2010; accepted October 2010