# Adapting Wikification to Cultural Heritage

**Samuel Fernando** and **Mark Stevenson**
Department of Computer Science
Regent Court
211 Portobello
Sheffield, S1 4DP
s.fernando@shef.ac.uk
m.stevenson@dcs.shef.ac.uk

## Abstract

Large numbers of cultural heritage items are now archived digitally along with accompanying metadata and are available to anyone with internet access. This information could be enriched by adding links to resources that provide background information about the items. Techniques have been developed for automatically adding links to Wikipedia to text but the methods are general and not designed for use with cultural heritage data. This paper explores a range of methods for adapting a system for adding links to Wikipedia to cultural heritage items. The approaches make use of the structure of Wikipedia, including the category hierarchy. It is found that an approach that makes use of Wikipedia's link structure can be used to improve the quality of the Wikipedia links that are added.

## 1 Introduction

Cultural heritage (CH) items are now increasingly being digitised and stored online where they can be viewed by anyone with a web browser. These items are usually annotated with metadata which gives the title of the item, subject keywords, descriptions and so on. However such metadata can often be very limited, with some items having very little metadata at all. This paper examines methods to enrich such metadata with inline links to Wikipedia. These links allow users to find interesting background information on the items and related topics, and provides a richer experience especially where the metadata is limited. Additionally the links may also help to categorise and organise the collections using the Wikipedia category hierarchy.

CH items from Europeana[1] are used for the evaluation. Europeana is a large online aggregation of cultural heritage collections from across Europe. The WikiMiner software (Milne and Witten, 2008) is used to automatically enrich the Europeana items collections with Wikipedia links. Two methods are used to improve the quality of the links. The first makes use of the Wikipedia category hierarchy. Top-level categories of interest are selected and articles close to these categories are used as training data for WikiMiner. The second method uses existing links from Wikipedia as evidence to find useful links for the CH items.

## 2 Background

Mihalcea and Csomai (2007) first addressed the task of automatically adding inline Wikipedia links into text and coined the term Wikification for the process. Their procedure for wikification used two stages. The first stage was *detection*, which involved identifying the terms and phrases from which links should be made. The most accurate method for this was found to be using link probability, defined as the number of Wikipedia articles that use the term as an anchor, divided by the number of Wikipedia articles that mention it at all. The next stage, *disambiguation* ensure that the detected phrases link to the appropriate article. For example the term *plane* usually links to an article about fixed wing aircraft. However it sometimes points to a page describing the mathematical concept of a theoretical surface, or of the tool for flattening wooden surfaces. To find the correct destination a classifier is trained using features from the context. Although the quality of re-

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

*Proceedings of the 6th EACL Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities*, pages 101–106,
Avignon, France, 24 April 2012. ©2012 Association for Computational Linguistics

sults obtained is very good, a large amount of preprocessing is required, since the entire Wikipedia encyclopedia must be parsed.

Milne and Witten (2008) build upon this previous work with the WikiMiner program. 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 WikiMiner software is freely available[2], and has been used as the basis for the approaches presented here.

Recent work on named entity linking and wikification makes use of categories and link information (Bunescu and Pasca, 2006; Dakka and Cucerzan, 2008; Kulkarni et al., 2009). Wikification has also been applied to the medical domain (He et al., 2011). Wikipedia categories and links have been used previously to find the similarity between CH items (Grieser et al., 2011). The category retraining approach presented here differs in that it only makes use of the top-level categories.

## 3 Methods

Three approaches to improving the quality of Wikipedia links added by WikiMiner were developed. The first two make use of Wikipedia's category structure while the third uses the links between Wikipedia articles.

### 3.1 Wikipedia Categories

Almost all articles in Wikipedia are manually assigned to one or more categories. For example the page ALBERT EINSTEIN belongs to the categories Swiss physicists, German-language philosophers and several others. The category pages thus group together articles of interest. Furthermore, each category may itself be a sub-category of one or more categories. So for example Swiss physicists is a sub-category of the categories Swiss scientists, Physicists by nationality etc.

The categories give a general indication of the topic of the article and we assume that articles relevant to Cultural Heritage items are likely to be closely associated with certain categories.

### 3.2 Retraining using Categories

The first approach is to retrain WikiMiner using articles associated with particular categories. Three top-level categories manually judged to indicate articles that are relevant to cultural heritage were selected: Culture, Arts and Humanities. All articles within 2 links of these selected categories were found and used as training data for WikiMiner. (We also explored using different numbers of links but found that fewer than 2 links produced a very small number of articles while more than 2 generated very large numbers which would be prohibitively expensive for retraining.) The same approach was also tested with categories which are unlikely to be related to cultural heritage (Computers, Mathematics and Science) in order to test the effect of using different categories.

### 3.3 Filtering using Categories

This approach uses the category information to filter articles after WikiMiner has been run. Each article added by WikiMiner is examined and any which are more than a certain distance from a top-level category which has been identified as being relevant to cultural heritage is removed. The assumption behind this approach is that relevant articles are much more likely to be closely associated with these categories than ones which are not relevant.

### 3.4 Exploiting Wikipedia's Link Structure

The final method makes use of Wikipedia's link structure rather than the category hierarchy and is similar to the previous method since it filters the links added by WikiMiner to identify those which are relevant to a particular article.

The first stage is to run the item through WikiMiner to detect suitable links. This is done with 2 parameter settings, each returning a set of links. The aim of the first run is to find as many

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

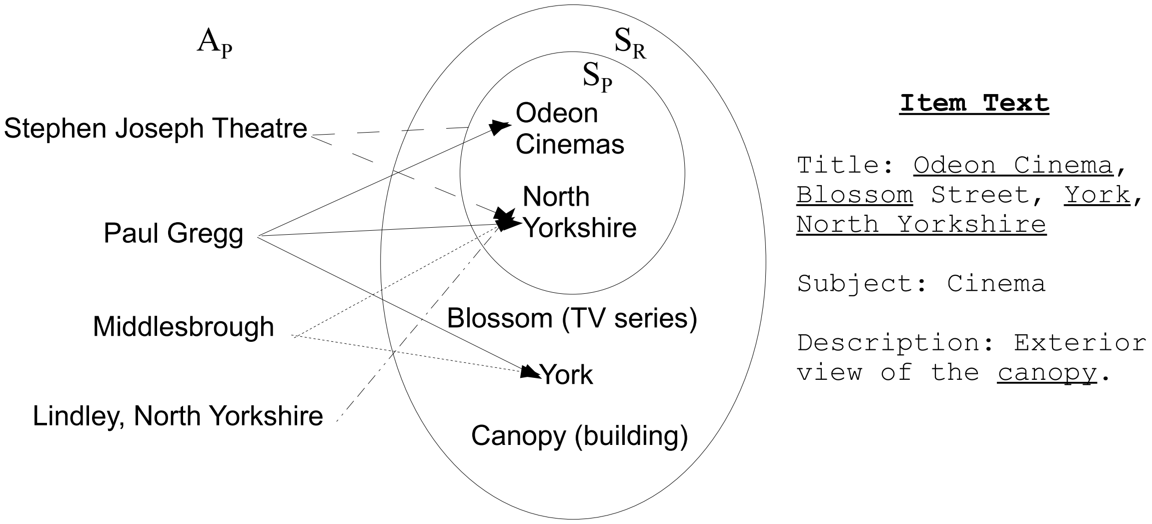

Figure 1: Example illustrating the method, where articles (on the left) which link to the high precision articles ($S_P$) are used to find good links in the high recall set ($S_R$).

potential links in the text as possible, for example by using a low confidence threshold. This give a set of links $S_R$ which is high recall (because most links are included), but low precision, since many incorrect or irrelevant links are also present. The aim of the second run is to find a smaller set of links which are likely to be of good quality, for example by setting a high confidence threshold. The resulting set $S_P$ is high precision but lower recall since good links may be discarded.

The result set of links is initialised with the high precision articles $R = S_P$. The aim is then to try to find additional good links within $S_R$. This is done by finding a list of articles $A_P$ which contain links to 1 or more of the articles in $S_P$. Let $O(a)$ be the set of outlinks from an article $a$. Each article in $A_P$ is then scored on how many links are shared with $S_P$:

$$\forall a \in A_P : score(a) = |O(a) \cap S_P| \qquad (1)$$

The $N$ top scoring articles in $A_P$ are then used to find further good links with within $S_R$. For each of these articles $a$:

$$R := R \cup (O(a) \cap S_R) \qquad (2)$$

Figure 1 gives an example illustrating how the method works on an Europeana item about an old Odeon Cinema in York. The article on Paul Gregg links to the articles in the $S_P$ set {Odeon Cinemas, North Yorkshire}. Since it also links to the York article in the $S_R$ set, the method takes this as evidence that York might also be a good article to link to, and so this would be added to the result set $R$.

## 4 Annotation

To evaluate the quality of the Wikipedia links, a sample of CH items was manually annotated. The sample of 21 items was randomly selected from Europeana. When run through WikiMiner with no probability threshold (i.e. including all possible links), a total of 366 potential links were identified. A further 16 links were manually added which the WikiMiner software had missed, giving a total of 381 links.

Web surveys were created to allow the annotators to judge the links. For each item in the survey users were presented with a picture of the item, the metadata text, and the set of possible links (with the anchor text identified). The annotators were then given a binary choice for each link to decide if it should be included or not.

Two separate surveys were taken by three fluent English speakers. The first was to determine if each link was correctly disambiguated within the context of the item (regardless of whether the link was useful or appropriate for that item). For each link the majority decision was used to judge if the link was indeed correct or not. Out of the 381 links, 70% were judged to be correct and 30% as incorrect. For 80% of the links the judgement was unanimous with all 3 annotators agreeing on the correctness of the links. The remaining 20% were 2-to-1 judgements. This gives an overall inter-annotator agreement of 93.4%.

The second survey was to determine which of the correct links were useful and appropriate for the corresponding items. As before each of the 21 items was presented to the annotators, but this time only with the 267 links that had been judged as correct within the previous survey. Again, three annotators completed the survey. Out of the 267 correct links, 49.8% were judged to be useful/appropriate and 50.2% as not. For 67.7% of the links the judgement was unanimous. The remaining 32.2% were 2-1 judgements. This gives an inter-annotator agreement of 89.3%. The 133 links judged to be correct, useful and appropriate were then used as the gold standard to evaluate the automatic methods.

As an example, the links and judgements for the following text are shown in Table 1:

Title: Odeon Cinema, Blossom Street, York, North Yorkshire
Subject: Cinema
Description: Exterior view of the canopy.

| Link | Correct | Useful |
|---|---|---|
| Odeon Cinemas | Yes | Yes |
| Blossom (TV series) | No | N/A |
| York | Yes | Yes |
| North Yorkshire | Yes | No |
| Cinema | Yes | No |
| Canopy | Yes | Yes |

Table 1: Examples of links and judgements

## 5 Experiments

The methods from Section 3 were used to identify links in the items from Europeana. The results were evaluated against the gold standard manually annotated data that was described in Section 4. For all experiments the standard metrics of precision, recall and F-measure are used to measure the performance of the methods.

Milne and Witten (2008) noted that training using articles with a similar length and link density to the target documents can improve WikiMiner's performance. The descriptions associated with Europeana items are relatively short so further experiments were carried out in which WikiMiner was retrained with different sets of articles. The best results were obtained using a set of articles between 100 and 500 words that contained a minimum of five links to other articles. (Results for experiments comparing other configurations are not reported here for brevity.) Table 2 shows results obtained using the default model, when WikiMiner is run 'off the shelf', and when it has been retrained. These results demonstrate that retraining WikiMiner improves performance. Precision improves to over 50% and, although there is a drop in recall, F-measure is also higher. Results using the retrained model are used as a baseline against which alternative approaches are compared.

| Model | P | R | F |
|---|---|---|---|
| Default | 34.0 | **91.7** | 49.6 |
| Retrained | **56.6** | 77.4 | **65.4** |

Table 2: Results obtained using WikiMiner using default model and after retraining

### 5.1 Category retraining

The category retraining approach (Section 3.2) was applied using all articles within two links of selected categories as training data for WikiMiner. The results are shown in Table 3 and show that precision is improved over the baseline for all categories. However the results do not fit the hypothesis, with Science giving the best F-measure overall, a statistically significant improvement over the baseline ($p < 0.05$, t-test). This may be for various reasons. Firstly the category hierarchy in Wikipedia is often messy with articles assigned to many different categories, and each category can contain a diverse sets of articles which may not be very useful. Secondly it may be that the topics of the articles are not so important for the training, but rather factors like the length of the articles and the link densities. However it is interesting that using articles close to the top level categories does appear to improve performance.

| Method | P | R | F |
|---|---|---|---|
| Baseline | 56.6 | **77.4** | 65.4 |
| Culture | 65.5 | 71.4 | 68.3 |
| Arts | 69.6 | 65.4 | 67.4 |
| Humanities | 71.9 | 65.4 | 68.5 |
| Mathematics | 72.9 | 58.6 | 65.0 |
| Science | 72.4 | 69.1 | **70.8** |
| Computers | **76.7** | 59.4 | 66.9 |

Table 3: Retraining using top level categories.

### 5.2 Category filtering

The category filtering approach (Section 3.3) was applied. Articles within a distance of 1 to 4 links from selected top level categories are kept and all others are discarded. The following combinations of categories were used: C (Culture), CHA (Culture, Humanities and Arts), and CHAGSE (Culture, Humanities, Arts, Geography, Society and Education).

Results are shown in Table 4 and are surprisingly low. Both precision and recall drop significantly when category filtering is applied. This may be because the articles within categories are often very diverse and do not capture many of the possible topics found within cultural heritage items.

| Method | Precision | Recall | F |
|---|---|---|---|
| Baseline | **56.6** | **77.4** | **65.4** |
| C | 35.1 | 19.5 | 25.1 |
| CHA | 27.4 | 27.8 | 27.6 |
| CHAGSE | 24.5 | 34.6 | 28.7 |

Table 4: Filtering using top level categories.

### 5.3 Using Wikipedia links

The final experiment explores the link filtering approach described in Section 3.4. The high precision $S_P$ set is chosen to be those returned by the retrained WikiMiner model ("Retrained" in Table 2) while the high recall $S_R$ set is the default model ("Default" in Table 2). Experiments were performed varying $N$, the number of top scoring articles used (using the score metric defined in Equation 1).

| No. of similar articles | P | R | F |
|---|---|---|---|
| Baseline | 56.6 | **77.4** | 65.4 |
| 1 | **74.0** | 53.4 | 62.0 |
| 2 | 70.7 | 61.7 | 65.9 |
| 3 | 68.5 | 63.9 | 66.1 |
| 4 | 67.4 | 68.4 | 67.9 |
| 5 | 66.9 | 69.9 | **68.4** |
| 6 | 66.2 | 70.6 | **68.4** |
| 7 | 66.2 | 70.7 | **68.4** |
| 8 | 65.5 | 71.4 | 68.3 |
| 9 | 65.1 | 71.4 | 68.1 |
| 10 | 63.9 | **72.9** | 68.1 |

Table 5: Filtering using Wikipedia's link structure

The results are shown in Table 5 and show a clear improvement in precision for N from 1 to 10. The F-measure peaks when 5-7 related articles are used. The improvement in the F-measure over the baseline is statistically significant ($p < 0.05$ t-test).

## 6 Conclusions and future work

This paper explores a variety of methods for improving the quality of Wikipedia links added by the WikiMiner software when applied to the cultural heritage domain. Approaches that make use of the Wikipedia category hierarchy and link structure were compared and evaluated using a data set of manual judgements created for this study.

The approaches based on the category hierarchy appeared to be less promising than those which used the link structure. Improvements were obtained by retraining WikiMiner using articles associated with particular categories. However the results were unexpected, with categories such as Science giving better performance as training data than categories such as Culture or Arts. Although a higher score was obtained using this method than the link approach, this may be due to factors such as document length and link density rather than the topic of the articles.

Results obtained using a novel method based on existing links within Wikipedia suggest this approach is promising. The method is fully unsupervised so it can be easily applied to domains other than cultural heritage.

Information from both categories and links could be combined in a similar way to that suggested by Grieser et al. (2011). Enriching cultural heritage data with Wikipedia links should improve the experience for users while they browse the data. In addition the links themselves may be useful to categorise, cluster and find similar items. Further work will investigate these possibilities.

## Acknowledgments

The research leading to these results was carried out as part of the PATHS project (`http://paths-project.eu`) funded by the European Community's Seventh Framework Programme (FP7/2007-2013) under grant agreement no. 270082.

## References

Razvan Bunescu and Marius Pasca. 2006. Using Encyclopedic Knowledge for Named Entity Disambiguation. In *Proceedings of European Chapter of the Association of Computational Linguistics (EACL)*, volume 6, pages 9–16.

Wisam Dakka and Silviu Cucerzan. 2008. Augmenting Wikipedia with Named Entity Tags. In *Proceedings of The Third International Joint Conference on Natural Language Processing (IJCNLP)*.

Karl Grieser, Timothy Baldwin, Fabian Bohnert, and Liz Sonenberg. 2011. Using Ontological and Document Similarity to Estimate Museum Exhibit Relatedness. *Journal on Computing and Cultural Heritage (JOCCH)*, 3(3):10.

Jiyin He, Maarten de Rijke, Maarten de Rijke, Rob van Ommering, and Yuechen Qian. 2011. Generating Links to Background Knowledge: A Case Study Using Narrative Radiology Reports. In *20th ACM Conference on Information and Knowledge Management (CIKM)*, pages 1867–1876, Glasgow.

Sayali Kulkarni, Amit Singh, Ganesh Ramakrishnan, and Soumen Chakrabarti. 2009. Collective Annotation of Wikipedia Entities in Web Text. In *Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 457–466.

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

Rada Mihalcea and Andras Csomai. 2007. Wikify!: Linking Documents to Encyclopedic Knowledge. In *ACM Sixteenth Conference on Information and Knowledge Management (CIKM)*, volume 7, pages 233–242.

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