The Sheffield and Basque Country Universities Entry to CHiC: using Random Walks and Similarity to access Cultural Heritage

Eneko Agirre[1], Paul Clough[2], Samuel Fernando[2], Mark Hall[2], Arantxa Otegi[1], and Mark Stevenson[2]

[1] University of the Basque Country, UPV/EHU
{e.agirre,arantza.otegi}@ehu.es
[2] University of Sheffield, Western Bank, Sheffield, S10 2TN, UK
{p.clough, s.fernando, m.mhall, m.stevenson}@sheffield.ac.uk

Abstract. The Cultural Heritage in CLEF 2012 (CHiC) pilot evaluation included these tasks: ad-hoc retrieval, semantic enrichment and variability tasks. At CHiC 2012, the University of Sheffield and the University of the Basque Country submitted a joint entry, attempting the three English monolingual tasks.
For the ad-hoc task, the baseline approach used the Indri Search engine. Query expansion approaches used random walks using Personalised Page Rank over graphs constructed from Wikipedia and WordNet, and also by finding similar articles within Wikipedia. For the semantic enrichment task, random walks using Personalised Page Rank were again used. Additionally links to Wikipedia were added and further approaches used this information to find enrichment terms. Finally for the variability task, TF-IDF scores were calculated from text and meta-data fields. The final results were selected using MMR (Maximal Marginal Relevance) and cosine similarity.

Keywords: Personalised PageRank, Random Walks, Information Retrieval, Wikipedia, WordNet, Knowledge Bases, Clustering, Maximal Marginal Relevance

1 Introduction

The Cultural Heritage in CLEF 2012 (CHiC) pilot evaluation proposed these tasks: ad-hoc retrieval, semantic enrichment and variability tasks.

The University of Sheffield and the University of the Basque Country submitted a joint entry, attempting the three English monolingual tasks.

2 Ad-hoc Retrieval Task

This task is a standard ad-hoc retrieval task, which measures information retrieval effectiveness with respect to user input in the form of queries. The topics

are based on real Europeana[3] query logs and the documents to be retrieved are metadata records of Europeana objects.

We participated in the English monolingual subtask and submitted 4 different runs: one baseline run, and other 3 runs applying query expansion.

For all our approaches, we used Indri search engine [1], which is a part of the open-source Lemur toolkit[4]. We indexed the title, subject and description fields of the Europeana objects. The Porter stemmer was used.

2.1 Baseline Approach

Our baseline approach is the default query likelihood language modeling method implemented in the Indri search engine. We chose the Dirichlet smoothing method, with the parameter μ set to the value 100. We refer to this approach as NOEXP run.

2.2 Query Expansion Approaches

Our query expansion retrieval model runs queries which contain the original terms of the query and the expansion terms. Documents are ranked by their probability of generating the whole expanded query (Q_{RQE}), which is given by:

$$P_{RQE}(Q_{RQE} \mid \Theta_D) = P(Q \mid \Theta_D)^w P(Q' \mid \Theta_D)^{1-w} \quad (1)$$

where w is the weight given to the original query and Q' is the expansion of query Q. The query likelihood probability ($P(Q \mid \Theta_D)$) is the one used for the baseline approach. Details about the probability of generating the expansion terms ($P(Q' \mid \Theta_D)$) are omitted here, please, refer to [5].

As we did not have training data, we fixed the parameters to the optimum values in other previous experiments ($w = 0.7$).

We use two different approaches to obtain the expansion terms.

Using Random Walks. The query expansion algorithm based on random walks over the graph representation of concepts and relations in a knowledge base to obtain concepts related to the queries.

We have use this approach for two different runs. The difference between these two runs is the knowledge-base used. We have used Wikipedia for one run (EXP_UKB_WIKI10 is the identifier for this run), and WordNet [2] for the other one (EXP_UKB_WN100). In order to obtain the graph structure of Wikipedia, we simply treat the articles as vertices, and the links between articles as the edges. We represent WordNet as a graph as follows: graph nodes represent WordNet concepts (synsets) and dictionary words; relations among synsets are represented by undirected edges; and dictionary words are linked to the synsets associated to them by directed edges. We used WordNet version 3.0, with all

[3] `http://www.europeana.eu`
[4] `http://www.lemurproject.org`

relations provided, including the gloss relations. This was the setting obtaining the best results in a word similarity dataset as reported by [3].

Given a query and the graph-based representation of Wikipedia or WordNet, we obtain a ranked list of related concepts as follows:

1. We first pre-process the query to obtain the lemmas and parts of speech of the open category words.
2. We then assign a uniform probability distribution to the terms found in the query. The rest of nodes are initialized to zero.
3. We compute personalized PageRank [4] over the graph, using the previous distribution as the reset distribution, and producing a probability distribution over WordNet concepts The higher the probability for a concept, the more related it is to the given document.

Basically, personalized PageRank is computed by modifying the random jump distribution vector in the traditional PageRank equation. In our case, we concentrate all probability mass in the concepts corresponding to the words in the query.

Let G be a graph with N vertices $v_1, \ldots, v_N$ and d_i be the outdegree of node i; let M be a $N \times N$ transition probability matrix, where $M_{ji} = \frac{1}{d_i}$ if a link from i to j exists, and zero otherwise. Then, the calculation of the *PageRank vector* $\mathbf{Pr}$ over G is equivalent to resolving Equation (2).

$$\mathbf{Pr} = cM\mathbf{Pr} + (1 - c)\mathbf{v} \tag{2}$$

In the equation, $\mathbf{v}$ is a $N \times 1$ vector and c is the so called *damping factor*, a scalar value between 0 and 1. The first term of the sum on the equation models the voting scheme described in the beginning of the section. The second term represents, loosely speaking, the probability of a surfer randomly jumping to any node, e.g. without following any paths on the graph. The damping factor, usually set in the [0.85..0.95] range, models the way in which these two terms are combined at each step.

The second term on Eq. (2) can also be seen as a smoothing factor that makes any graph fulfill the property of being aperiodic and irreducible, and thus guarantees that PageRank calculation converges to a unique stationary distribution.

In the traditional PageRank formulation the vector $\mathbf{v}$ is a stochastic normalized vector whose element values are all $\frac{1}{N}$, thus assigning equal probabilities to all nodes in the graph in case of random jumps. In the case of personalized PageRank as used here, $\mathbf{v}$ is initialized with uniform probabilities for the terms in the document, and 0 for the rest of terms.

PageRank is actually calculated by applying an iterative algorithm which computes Eq. (2) successively until a fixed number of iterations are executed. In our case, we used a publicly available implementation[5].

In order to select the expansion terms, we choose the top N highest scoring concepts. When using Wikipedia, the first 10 concepts are used as expansion

[5] `http://ixa2.si.ehu.es/ukb/`

terms. In the case of WordNet, we get all the words that lexicalize the first 100 concepts.

For instance, given a query like "*Esperanto*", this method based on Wikipedia suggests related terms or phrases like *L. L. Zamenhof*, *interlingua*, *international auxiliary language* and *constructed language*.

Using Wikipedia Similarities. For the second query expansion approach, we use the 10 concepts obtained by the WIKISIM approach for the Semantic Enrichment task (see Sec. 4). We refer to this query expansion approach as the EXP_SE_WIKISIM run.

2.3 Results

Method	MAP
EXP_UKB_WN100	51.61
NOEXP	51.48
EXP_SE_WIKISIM	50.96
EXP_UKB_WIKI10	50.64

Table 1. Mean average precision for all Ad-hoc experiments.

The approaches from our submission give the best results overall in comparison to the other submissions. The baseline NOEXP approach provides strong results. However the query expansion approaches give little improvement, or even slightly degrade performance.

3 Variability Task

The goal of this task is to present a list of 12 items that give a good overview [7] over what types of items are available for the given query. To achieve this we have investigated two methods for selecting the 12 items to display and two sources for the meta-data that the selection algorithms work on.

Indexing and searching of the collection was performed using Apache Solr[6]. By default only the *dc:title*, *dc:description*, and *dc:subject* fields were searched and all words in the query were required to appear in those fields. Basic fully automatic query expansion was used to achieve higher recall. For singular nouns the plural form was added as a search keyword, vice-versa for plural nouns the singular form was added. If a word looked like a year, then the *enrich:period_label* and *europeana:year* fields were also searched for that year. Additionally we used

[6] `http://lucene.apache.org/solr/`

a very small gazetteer of place-names to identify candidate toponyms which were then searched for in the *enrich:place_label* field.

For each of the query result documents we then calculated two TFIDF scores, one based on the textual description of the items (labelled TEXT, using *dc:title*, *dc:description*, and *dc:subject* fields), the other based on what we termed the meta-data facets (labelled FACET, using *dc:subject*, *europeana:dataProvider*, *enrich:place_label*, *europeana:year*, *europeana:type*, and *dcterms:medium* fields). For the textual descriptions the *dc:title* and *dc:description* were sentence and word tokenised, and then stop-worded using NLTK[7]. For all other fields the whole field content was used as a single token. Using the TFIDF scores the final 12 documents were then selected using either Maximal-Marginal-Relevance (MMR) [9] or cosine-similarity.

3.1 Approaches

Maximal-Marginal-Relevance The selection using MMR starts by clustering the documents using k-means. The number of clusters k is selected automatically [6]. The MMR algorithm then iterates over the resulting clusters, each time selecting a document from the cluster that is most dissimilar (using cosine similarity) to the documents that have already been selected. Additionally if a document's title is the same as a title in the list of previously selected documents, it is skipped, unless that would reduce the final number of documents to less than 12.

Cosine Similarity In the cosine similarity method we randomly select a document to be the first document. The remaining documents are then sorted by decreasing cosine similarity to the first document. From this ranking we then sample the final 12 documents at regular intervals.

3.2 Results

Method	MAP
CLUSTERFACETS	23.93
CLUSTERTEXT	23.13
SIMFACETS	22.59
SIMTEXT	21.85

Table 2. Mean average precision for all Variability experiments.

The results from our submission give the best results overall when compared to the other submissions. As the variability judgements have not yet been released, the quality of the results is hard to judge. However the results seem to

[7] `http://nltk.org/`

imply that the cosine similarity produces better results if there are about 12 topics in the results and the topics have roughly the same number of items each, whereas the MMR method works better if this condition does not hold. The most likely reason for this is that the regular sampling used in the cosine similarity method breaks if the topic distribution is heavily skewed, or there are only a few topics.

4 Semantic Enrichment Task

The goal of this task is to present a ranked list of at most 10 concepts (words or phrases) for each query. These concepts should semantically enrich the query and/or guess the information need or query intent of the user. The concepts can be extracted from the Europeana data that has been provided (internal) or make use of other resources such as the Web or Wikipedia (external).

4.1 Approaches

Random walks The concepts from the UKBWIKI and UKBWN runs are obtained using Wikipedia and WordNet, respectively, following the expansion strategy based on random walks explained in Section 2.2.

Wikipedia links Two approaches attempted to find useful terms by finding inline Wikipedia links within each of the items in the collection. So for example given the text:

```
Hiroshima peace lanterns at Leith 1985.
Leith; Ceremonies; Peace demonstrations; Eighties;
War in Japan; World War II keywords
```

Links might be added to the terms `Hiroshima, Leith` and `World War II`. The motivation behind this approach is that these added links will suggest useful keywords which co-occur with the query term and so could be used as semantic enrichments for the term.

The Wikipedia Miner software [8] was used to find the links. This software has been trained on a Wikipedia snapshot from 6th Jan 2011. The software attempts to learn from the way Wikipedia itself links to other articles. The main training features used are commonness and relatedness of anchor terms. The commonness feature measures how often a certain anchor text links to a particular article in the text (so for example 'tree' will link more often to the plant than to the more obscure computer science definition of tree). The relatedness computes how closely related the term is to others in the text that is being linked, and thus takes into account the context of the text.

The first approach used an IR engine to find items that contained the query term. The IR engine used was Apache Solr (as described in Section 3). All returned items were then run through Wikipedia Miner to markup Wikipedia

links. A confidence threshold of 0.2 was used to eliminate low confidence links. The links that most often co-occurred within the items were then returned as the semantic enrichments. So for example World War II occurred frequently as a link in items containing 'Hiroshima', and so was returned as a enrichment term. This method is referred to as QUERYLINKS.

The second approach used a slightly different method. Here a previously processed version of Europeana was used which had already been run through Wikipedia Miner to find links for all items. Then instead of searching for items containing the query term using the IR engine, items that contained a link to the query term were used instead. So for example instead of searching for 'hiroshima', only items that contained a link to 'Hiroshima' were used. As with the first approach the most frequently co-occuring links were returned as the enrichment terms. One problem with this method is that sometimes very few or no items were found which contained a link to the query term. So as a fallback, a different method was used which made reference only to Wikipedia. The method found articles that were most similar to the query term, by examining the in and outlinks to and from the article and returning the article titles which contained the highest proportion of these in/outlinks. This method is referred to as WIKISIM.

4.2 Results

Method	MAP
UKBWIKI	29.05
UKBWN	20.70
WIKISIM	19.29
QUERYLINKS	16.61

Table 3. Mean average precision for all Semantic Enrichment experiments.

The UKBWIKI was the strongest run from our submission, showing the effectiveness of the random walk approach. The difference between the UKBWIKI and UKBWN results shows that the richness of Wikipedia provides a substantial benefit over WordNet. The WIKISIM and QUERYLINKS approach come around the middle of the table, with WIKISIM faring substantially better. This would seem to show that returning similar articles based on links seems to be quite effective for this task.

5 Conclusions

For the ad-hoc retrieval task, a strong baseline approach achieved better results than all the other submissions. The query expansion approaches presented here did not improve substantially on this performance.

Our submission for the variability task achieved the best results overall using text and facet similarity calculations to select the items to return. The MMR clustering approach proved slightly more effective than the cosine similarity measure.

For the semantic enrichment task the random walk approach proved effective, giving the 3rd and 6th overall highest precisions. The richer graph produced from Wikipedia proved to give substantially better results than using WordNet. The Wikipedia link approach gave results about midway in the table. Finding similar articles (in the WIKISIM run) proved to be slightly more effective than using the links alone (QUERYLINKS).

Acknowledgements

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 and KNOW2 project (TIN2009-14715-C04-01).

References

1. Strohman, T., Metzler, D., Turtle, H., Croft, W.B.: Indri: a language-model based search engine for complex queries. Technical report, in Proceedings of the International Conference on Intelligent Analysis (2005)
2. Fellbaum, C., editor: WordNet: An Electronic Lexical Database and Some of its Applications. MIT Press, Cambridge, Mass (1998)
3. Agirre, E., Soroa, A., Alfonseca, E., Hall, K., Kravalova, J., Pasca, M.: A Study on Similarity and Relatedness Using Distributional and WordNet-based Approaches. In Proc. of NAACL, Boulder, USA (2009)
4. Haveliwala, H: Topic-sensitive pagerank. In WWW 02, pages 517526, New York, NY, USA (2002).
5. Otegi, A., Arregi, X., Agirre, E.:Query Expansion for IR using Knowledge-Based Relatedness. Proceedings of 5th International Joint Conference on Natural Language Processing, pages 1467–1471 (2011)
6. Sugar, C. A. and James, G.M.: Finding the Number of Clusters in a Dataset. Journal of the American Statistical Association 98 (463), pages 750–763 (2003)
7. Hornbaek, K. and Hertzum, M.: The notion of overview in information visualization. International Journal of Human-Computer Studies 69, pages 509–525 (2011)
8. Milne, D. and Witten, I.: Learning to Link with Wikipedia. In Proceedings of the 17th ACM conference on Information and Knowledge Management, pages 509–518 (2008)
9. Carbonell, J. and Goldstein, J.: The use of MMR, diversity-based reranking for reordering documents and producing summaries. In Proceedings of the 21st Annual international ACM SIGIR Conference on Research and Development in information Retrieval. SIGIR '98. ACM, New York, NY, pages 335–336 (1998)