DiscOU: A Flexible Discovery Engine for Open Educational Resources Using Semantic Indexing and Relationship Summaries

Mathieu d'Aquin, Carlo Allocca and Trevor Collins
Knowledge Media Institute, The Open University, Milton Keynes, UK
{m.daquin, c.allocca, t.d.collins}@open.ac.uk

Abstract. We demonstrate the DiscOU engine implementing a resource discovery approach where the textual components of open educational resources are automatically annotated with relevant entities (using a named entity recognition system), so that these rich annotations can be searched by similarity, based on existing resources of interest.

1 Introduction / Motivation

There is a growing base of open educational content being made available online. At the Open University, this currently includes 650 units of course material on OpenLearn and 3,800 audio and video podcasts[1]. With such content available, discoverability of educational resources becomes a major challenge. The exposure of the metadata for such resources as linked data (see data.open.ac.uk and [1]) is expected to make these resources more directly addressable, together with their general description and the subjects they are covering (see e.g. [2]). Accordingly, linked data is increasingly being adopted in open and distance learning scenarios where discoverability is a main requirement (see [3]). However, relying purely on metadata requires either to stay at a high level of description of the content of resources (through the general topics being covered) or to richly annotate these resources with all the dimensions relevant to their content. Another common approach is therefore to search by similarity based on existing resources of interest (i.e. finding things that are "more like this"). This is however generally limited to the comparison of the textual components of the resources, with obvious limitations, as similarity between texts does not necessarily reflect a useful relationship between resources in a discovery scenario.

Here, we demonstrate an engine implementing a hybrid approach, where the textual components of open educational resources are automatically annotated with relevant entities (using a named entity recognition system), so that these rich annotations can be searched by similarity. This allows us to discover resources based on relationships that are not necessarily explicitly described in their metadata, and to characterise semantically these relationships based on shared entities. This also provides us with a more flexible workflow, compared to typical recommendation engines, where the user can act upon the search for resources, through customising the semantic annotations realised prior to similarity search. We demonstrate a prototype application of the developed services to discover open educational content from the Open University, based on the content of programmes broadcasted by the BBC.

[1] see http://podcast.open.ac.uk and http://openlearn.open.ac.uk

2 The DiscOU Approach

Figure 1 summarises the architecture of the DiscOU system, which describes the workflow implemented in four RESTful services[2] for 1- extracting semantic entities from an online resource, 2- indexing these entities, 3- searching by similarity in the index and 4- summarising the relationships between resources.

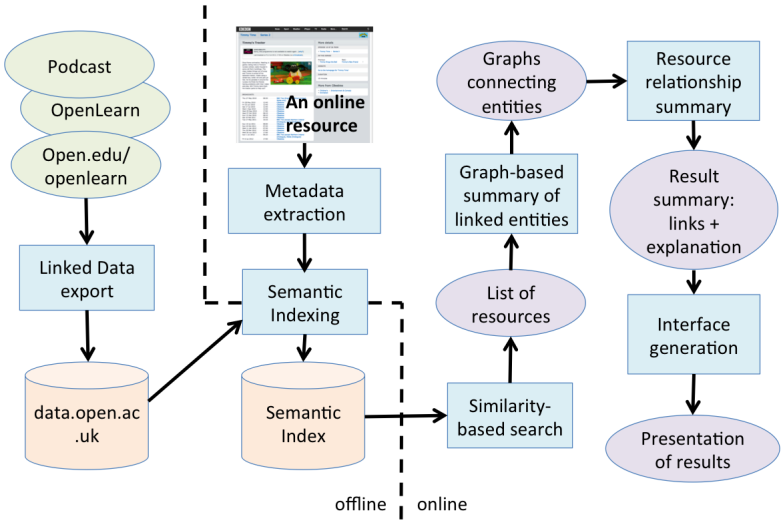

Fig. 1. Overview of the architecture of the DiscOU system.

Semantically Indexing Online Resources. As described in the previous section, the main idea behind DiscOU is to take a hybrid approach where resources can be searched "by similarity" with another existing resource, but where the comparison of resources is based on rich semantic annotations. To generate such rich semantic annotations, we make use of a named entity recognition system, namely DBpedia Spotlight [4]. Textual components are first extracted from the resources to be indexed, based on their description on data.open.ac.uk (using the metadata directly for the title and abstract, and links to the textual content, as online documents for OpenLearn units and PDFs of transcripts for podcasts). The online service provided by DBpedia Spotlight is then used to obtain a list of DBpedia entities with confidence/relevance scores for each of these components.

To index these semantic descriptions, we use the Lucene open source search engine library[3]. Lucene is however designed to index documents and texts and is based on term-occurrence measures for searching and ranking results (i.e., TF.IDF). The indexing of semantic annotations is therefore realised in such a way that these mechanisms can be used to obtain relevant results when searching on the basis of semantic entities rather than of text. This is achieved simply by

[2] see http://discou.info
[3] http://lucene.apache.org/core/

transforming the relevance score provided by DBpedia Spotlight into a number of occurrences for the entity, therefore repeating the mention of an entity in the index of a given resource depending on its relevance for the resource. In this way, when searching based on semantic entities, Lucene should return in priority resources for which these entities are highly relevant.

Searching by Similarity. Lucene provides the base technique to search by similarity through a mechanism called "MoreLikeThis". This mechanism takes as input an indexed document and generates a query that is expected to return other resources having similar indexes. We apply this mechanism through first indexing the external resource used as starting point for the discovery process (in the next section, we use BBC programme webpages) using the same process as described above. Because of the way the index is constructed, resources are returned that share a large part of their semantic annotation.

Summarising Relationships Between Resources. One major advantage of our approach is that the similarity relationship between resources being discovered and the original 'query resource' is characterised by the semantic entities shared in their content. Depending on the richness of the considered content however, such lists of shared entities can be too large to be useful summaries of this relationship. To tackle this issue, we developed a mechanism to summarise lists of DPpedia entities. It uses DBpedia links between entities in a list (using a local index of DBpedia, optimised for this specific task) to generate a set of connected graphs. Each of these graphs is expected to represent one major topic of the resources being considered. We therefore select the one which contains the most entities with the highest relevance and, within this graph, the entity that appears to be the most connected and the most relevant.

3 Demonstrator: Finding Open University Content Based on BBC programmes

We implemented a demonstrator using the above mechanisms for a scenario in which a user, having found a BBC programme interesting, wants to obtain links to open educational resources to learn more about the topics covered by the programme (see Figure 2). The interface of the demonstrator is implemented in Javascript, using a bookmarklet to trigger the search. In other terms, being on a BBC programme page (in the example Figure 2, "The Secret Life of Chaos"[4]), the user can click on the DiscOU bookmark to make appear the results of searching for similar resources in Open University content.

This is realised by extracting textual content from the BBC programme page (out of its RDF description on the BBC website), running the semantic indexing service on this content and searching by similarity. The results show the titles and descriptions of the retrieved resources (obtained using the SPARQL endpoint of data.open.ac.uk) as well as the summary of the relationship between each resource and the BBC programme (here, mostly that they are about Chaos Theory and Economy). One interesting aspect of this demonstrator is that the

[4] http://www.bbc.co.uk/programmes/b00pv1c3

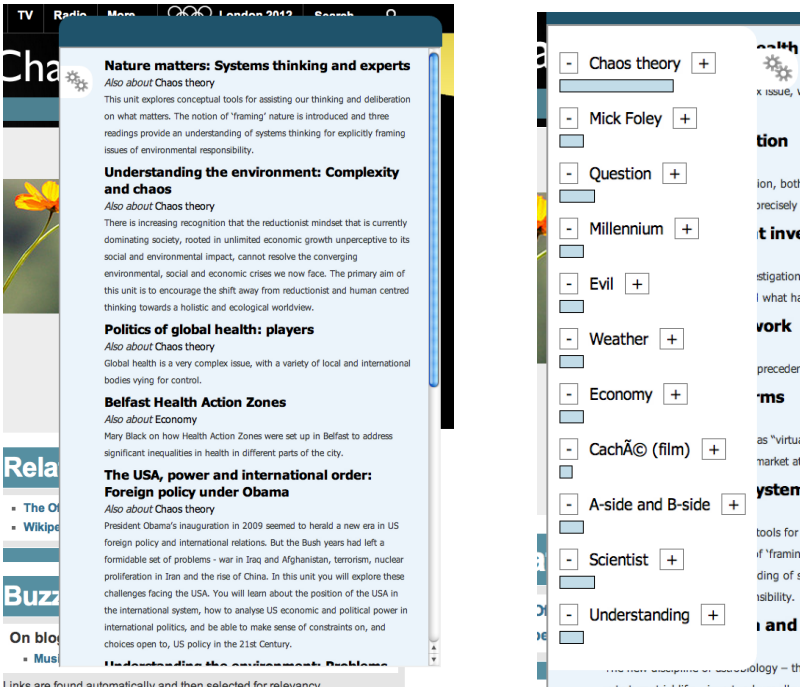

Fig. 2. Default results obtained with the BBC programme "The Secret Life of Chaos" (left) and interface to customise the semantic annotations for this programme (right).

user can customise the 'query' by changing the weights of the entities extracted as semantic annotations for the BBC programme (see right part of Figure 2). Once the weights are customised, the search is triggered again, showing results that are related to the personalised semantic annotations of the BBC programme.

While the results obtained are not always relevant, the fact that some level of explanation is provided together with the ability to refine the automatically generated 'query' makes the issue of incorrect results less critical. It is worth mentioning in particular that only a very small part of the system is specific to BBC programmes. The engine is used by the demonstrator as a set of RESTful services, with its functionalities being highly reusable in other scenarios.

References

1. d'Aquin, M.: Putting linked data to use in a large higher-education organisation. In: Interacting with Linked Data workshop. (2012)
2. Heath, T., Singer, R., Shabir, N., Clarke, C., Leavesley, J.: Assembling and applying an education graph based on learning resources in universities. In: Linked Learning (LILE) Workshop. (2012)
3. d'Aquin, M.: Linked data for open and distance learning. Commonwealth of Learning report (2012)
4. Mendes, P., Jakob, M., García-Silva, A., Bizer, C.: Dbpedia spotlight: shedding light on the web of documents. In: International Conference on Semantic Systems. (2011)