Smart collections: can artificial intelligence tools and techniques assist with discovering, evaluating and tagging digital learning resources?

Richard Leibbrandt, Dongqiang Yang, Darius Pfitzner and David Powers

School of Computer Science, Engineering and Mathematics, Flinders University GPO Box 2100, Adelaide SA 5001, Australia
http://csem.flinders.edu.au Email: richard.leibbrandt@flinders.edu.au

Pru Mitchell, Sarah Hayman and Helen Eddy

Education Network Australia (edna), Education Services Australia, 182 Fullarton Road, Dulwich, SA 5065, Australia
http://ww.edna.edu.au Email: pru.mitchell@esa.edu.au

Abstract

This paper reports on a joint proof of concept project undertaken by researchers from the Flinders University Artificial Intelligence Laboratory in partnership with information managers from the Education Network Australia (edna) team at Education Services Australia to address the question of whether artificial intelligence techniques could be employed to help with creation and consistency of learning resource metadata and improve the efficiency of digital collection workflows? The results show some success with automated subject categorisation on a small sample, and the researchers conclude that automated classification based on artificial intelligence is useful as a means of supplementing and assisting human classification, but is not at this stage a replacement for human classification of educational resources.

Introduction

Digital learning resources represent a significant investment and priority for 21st century educators, but the quantity and nature of digital content means that manual identification and cataloguing of appropriate resources is time-consuming and can result in inconsistent metadata which is not readily shared between systems.

This paper describes the Smart Collections project, a joint proof of concept project undertaken by researchers from the Flinders University Artificial Intelligence Laboratory in partnership with information managers from the Education Network Australia (edna) team at Education Services Australia during 2008-2009. As a cross-disciplinary, cross-institutional team the challenge was to see what could happen when artificial intelligence met the challenge of education metadata creation.

Background

Education Network Australia (edna) was established in 1996 as an online service to support and promote the benefits of technology for education and training in Australia. It is a collaborative information service which is funded and developed in partnerships with the Australian education and training community. edna is an aggregator service. It investigates, compiles, filters, evaluates, annotates and consolidates quality information and provides access to online resources news, networks, events, projects and research for education. The project supports a set of websites, collaborative workspaces, discussion lists, professional networking services and xml-based information services which are used by stakeholders on their own websites, portals, RSS readers and handheld devices. Content syndication via RSS, federated search and federated security are key features of the edna project. The edna metadata profile is based on the Dublin Core Metadata Initiative (1998) and the edna metadata standard v1.1 (Education Network Australia, 2001). A team of information officers at edna is responsible for building and maintaining a large collection of web-based learning resources with associated metadata.

The Flinders University Artificial Intelligence and Knowledge Discovery laboratory focuses on issues relating to the discovery, modelling, interpretation and use of information and knowledge. While the

 Diversity Challenge Resilience: School Libraries in Action Proceedings of the 12th Biennial School Library Association of Queensland, the 39th International Association of School Librarianship Annual Conference incorporating the 14th International Forum on Research in School Librarianship, Brisbane QLD Australia, 27 September – 1 October 2010.

emphasis is on applied research and the group strives to develop solutions to real-world data-rich problems, the laboratory is also interested in fundamental technologies, particularly in artificial intelligence and information theory.

The Smart Collections project grew out of a previous proof of concept project initiated at Education Network Australia (edna) as part of its 2007 innovations contract, called edna semantic collections (ESC). The overall vision of the ESC project was "to enhance the quality, size and diversity of the edna digital resource collection through user engagement, automation and improved metadata tools" (Hayman & Lothian, 2009, p. 119). This was a response to several changes in the educational resource collection landscape including the rapidly increasing quantity and diversity of digital content and reduced staff budgets, which meant it was no longer possible to manually identify and catalogue resources to the level described in the edna collection policy (Education Network Australia, 2007). At the same time there was increasing value placed on creating consistent metadata to facilitate sharing of resources between education systems.

The user engagement component of this vision sought to capitalise on the changes in online communication made possible with the advent of Web 2.0 and involved the harvesting of user-generated metadata from the social bookmarking service, *Delicious*, then evaluating its relevance to the collection, and how this metadata could be integrated into edna workflows (Lothian, 2007).

The second part of the strategy involved consideration of better metadata tools, in particular automated metadata processes and extraction. It also sought to capitalise on the beginnings of semantic annotation by applying "additional information that identifies or defines a concept in a semantic model in order to describe part of that document" (Farrell & Lausen, 2007). Researchers from the Flinders University Artificial Intelligence Laboratory were at a show and tell presentation for the edna ESC project and from this stemmed the partnership formed to undertake this Smart Collections project. The goal for the joint Flinders University – Education Network Australia research team was to find ways in which artificial intelligence techniques could be employed to help with collection workflow and to provide greater consistency of learning resource metadata.

In the edna workflow at that time, metadata records were created using online forms within the edna-customised installation of the DSpace open source repository platform. The human metadata creator entered the url of an online resource, the system sought to 'scrape' any Dublin Core metadata available from the document at that url, and the human filled in all the fields left empty that were relevant to the resource being catalogued according to the guidelines provided in the edna resources metadata profile (Education Network Australia, 2006) or the events metadata profile (Education.au, 2007). A small number of vocabulary tools were available within the DSpace form for the selection of terms from controlled lists in such elements as subject, audience, type, format, coverage, rights and edna category. This was a very manual process and made more complex by the use of multiple thesauri to describe resources for multiple sectors of education served by edna, namely early childhood, schools, vocational education and training, higher education, adult and community education.

In order to keep up with the rapidly growing body of educational resources online and continue to provide edna users with timely information services based on relevant and appropriately targeted information, Education Services Australia and the Artificial Intelligence Laboratory explored the use of modern artificial intelligence and information retrieval techniques in search of solutions that might be available to partially automate and accelerate the growth of the edna repository and increase the diversity of content types, whilst maintaining high quality standardised metadata.

Research questions

Prior to investigating strategies for improving collection and metadata processes it was important to consider what decisions metadata creators were required to make in the digital collection building workflow, and how important each of these decisions was to the end-user's discovery and assessment of relevance. Evaluating the relative cost of each of these decisions in terms of time, expertise and effort was useful in determining priorities for the project.

 Diversity Challenge Resilience: School Libraries in Action Proceedings of the 12th Biennial School Library Association of Queensland, the 39th International Association of School Librarianship Annual Conference incorporating the 14th International Forum on Research in School Librarianship, Brisbane QLD Australia, 27 September – 1 October 2010.

The primary research question for this project was then stated as 'which steps in the metadata creation workflow, that is which of the decisions made by human metadata creators, could artificial intelligence systems be trained to implement'? Supplementary questions flowing from this primary research question included what systems could be developed to improve the metadata workflow, what level of effectiveness could be realised and what benefits could adoption of these systems bring to digital collection building?

Our aim was to optimise the data collection cycle by augmenting manual collection processes with automated ones or replacing repetitive human processes where possible. The team considered there could be several potential benefits from this investigation.

1. Improved efficiency if an automated system could provide classification suggestions acceptable to metadata creators,
2. Improved user experience of subject and category-based searches if an automated system could deliver more relevant results through quicker, more consistent application of thesaurus structures and terms, and
3. Improved integration of user contributed resources if an automated system could map user tags to controlled thesaurus terms.

Literature

While proof of concept project methodology doesn't traditionally require a literature review, it is useful to compare the project's question and methodology with other similar projects or research activity. Other organisations identified the same issues with scalability and quality of metadata creation that formed the major impetus for this project. Several reports of investigation into automated metadata techniques were reviewed, as well as previous studies of the artificial intelligence techniques used for this project, such as text classification.

Metadata

In Australia, the University of Tasmania, State Library of Tasmania, Department of Education, and TAFE Tasmania (Rowlands, 2004) investigated the impact on searcher satisfaction of using a minimal metadata set to describe learning objects rather than the usual larger set of metadata elements. In the course of this study the researchers found that "educators are most interested in subject relevance when searching for materials… metadata elements such as level, creator, sector, type and format were not missed by participants" (p. 13). This finding was pertinent to decisions about which metadata elements to make the priority for investigation.

Automated metadata as a possible solution to the issue of too many potential resources and too few human resources to catalogue them was found to be the focus of a number of international studies (Cardinaels, Meire & Duval, 2005; Greenberg, Spurgin & Crystal, 2006; Paynter, 2005). Researchers had studied a range of metadata types and automation techniques and as Paynter (2005) concluded, "the metadata evaluation tools have proved their value on numerous occasions, though we stress that they are imperfect surrogates for formal human evaluations" (p. 300). Ochoa & Duval (2007) found that "thanks to recent developments on automatic generation of metadata and interoperability between repositories, the production, management and consumption of metadata is vastly surpassing the human capacity to review or process this information" (p. 1). Their warning was that large scale automation of metadata will require new strategies of ensuring metadata quality.

Artificial intelligence

Artificial Intelligence involves the use of computer programs to solve problems for which there is no simple and straightforward process of arriving at a solution. Instead, these problems require the use of more complex forms of knowledge, for instance examples of previous problems and their solutions, or explicitly-stated rules and heuristics (provided by experts) which can be used to inform a decision but do not simply provide an answer.

Text Classification (TC) is a common task within artificial intelligence, and aims to thematically or semantically disambiguate natural language texts using one or multiple topical tags. These tags may be

 Diversity Challenge Resilience: School Libraries in Action Proceedings of the 12th Biennial School Library Association of Queensland, the 39th International Association of School Librarianship Annual Conference incorporating the 14th International Forum on Research in School Librarianship, Brisbane QLD Australia, 27 September – 1 October 2010.

hierarchically organized into web categories ordered from most general to most specific, for example from art to movies to home video, or may form a flat group of elements for example, positive, neutral, and negative for affect annotation. An example of a well-known classification hierarchy is the Dewey Decimal Classification System with three levels and a total of 1000 sections. The implementation of TC often relies on machine learning techniques, that is learning its own classification rules from examples, given the expensive cost of compiling disambiguation rules from experts' knowledge. Sebastiani (2002) provides a review of developments in TC.

Earlier research in Text Classification focused on adding semantic features to the representation of a document, that is, to augment the raw text with representation and then carefully training topic learners. A more recent approach is to calculate the semantic relatedness between terms, that is a quantitative measure of how similar or thematically related the two terms are in their meanings. Most work on semantic relatedness in text classification (Budanitsky & Hirst, 2006; Pedersen, Patwardhan & Michelizzi, 2004) has made use of *WordNet* (Fellbaum, 1998), a standard ontology of English words and the semantic relationships between them. *WordNet* (Princeton University, 2010) describes hierarchies of words in terms of both hyponymic (e.g. France is a country) and meronymic relationships (e.g. a dog has a tail) between them. Consequently, *WordNet* implicitly expresses how close two words are in meaning, by the number of links in the hierarchy that need to be traversed in order to move from one word to the other. Intuitively, a pair of words that are connected by only a few semantic links are more closely semantically related than a pair of more distant words. More recently, many researchers have also found *Wikipedia* (2010) to be an attractive alternative source of semantic information (Ponzetto & Strube, 2007; Turdakov & Velikhov, 2008).

It is possible, but computationally expensive, to make use of the entire text of a document as information that can help a TC program to assign the document to a particular category. Instead, an important subtask in TC is typically to extract from the text just a small number of keywords and key phrases that are particularly indicative of the topic of the document, and then to classify based on the keywords rather than the entire text. One of the most popular techniques for key phrase extraction is Text Frequency/Inverse Document Frequency or TFIDF (Ramos, 2003) which identifies those phrases that are most distinctive of a document (roughly speaking, the phrases that occur relatively often in the particular document and relatively seldom in other documents). However, the phrases identified by TFIDF are not always the same phrases that human readers would choose to describe the topic of a document (Pfitzner, Treharne & Powers, 2008). Hence, one of the challenges of research is to make progress in delineating a number of heuristics by which experts locate the key phrases in a text that allow them to decide on appropriate metadata topic/category terms.

Methodology

The Smart Collections project adopted an ongoing proof of concept methodology (Cotton, 2007) supporting investigations into several elements of the overall research questions listed above as time, resources and expertise permitted. It sought to apply advances made by the Flinders Artificial Intelligence Lab to the management of educational resource collections.

edna makes use of a number of different classification systems depending on the target audience and initial categorisation of a document. One of the key elements of this investigation revolved around automation of subject description. Metadata in edna provides subject access via several controlled vocabularies, some internally developed and maintained and others from external agencies. edna categories provide a broad classification hierarchy and are used to deliver the browse functionality in the edna website. Learning resources categories in edna are linked to subject or discipline names. In the school sector the categories reference curriculum strands. In the higher education sector edna categories reference the national research discipline classification schema, and in the vocational education and training sector the Australian industry categories are used.

For the purposes of the Smart Collections project investigations it was decided to focus on use of two vocabularies, edna categories and the Schools Online Thesaurus (ScOT). ScOT (Education Services Australia, 2010) provides a controlled vocabulary of terms used in Australian and New Zealand schools. It encompasses all subject areas as well as terms describing educational and administrative processes. ScOT

 Diversity Challenge Resilience: School Libraries in Action Proceedings of the 12th Biennial School Library Association of Queensland, the 39th International Association of School Librarianship Annual Conference incorporating the 14th International Forum on Research in School Librarianship, Brisbane QLD Australia, 27 September – 1 October 2010.

terms are used to describe school education resources in edna, particularly curriculum resources. They provide much more specific subject access than the broader edna categories.

In the text classification task for this project, we were interested in

1. mapping from full-text documents (ideally, using key phrases, but other forms of available metadata would also be appropriate) to several relevant subject keywords from a controlled vocabulary, in this case the Schools Online Thesaurus (ScOT); and
2. mapping from the full-text (again, using key phrases or metadata) to the relevant edna category (or curriculum learning area strand).

These two tasks can be viewed as text classification tasks, or more precisely topic classification tasks, as we were trying to assign a diverse set of English words and phrases to one or more topic labels taken from a restricted set of options.

One fairly standard solution to this kind of problem is to determine the statistical correspondence between individual pairs of items from the two data sets (for example between a particular key phrase and a particular controlled vocabulary term). This approach estimates the probability that a particular language item from the first set will co-occur with an item from the second set, given the number of times that the two items have in fact co-occurred in the past. Two problems with this approach in the current context are that:

1. it requires a large amount of historical data, and
2. it can only deal with words and phrases that have already occurred in edna resources in the past.

These problems stem from the fact that a statistical approach to language data is not able to understand the meaning of the words that it processes. In this project, we have begun to explore a more promising approach to text classification, which attempts to use information about how words are meaningfully related to each other, in order to decide which appropriate ScOT terms and edna categories should be assigned to a resource.

Three tasks were identified as potentially benefiting from automation, and separate experiments were designed to investigate elements of these three tasks. In experiment 1 keyword extraction was used to determine whether key phrases from the text could be used to predict the subject (dc.subject), audience (edna-audience) and education sector (edna-sector) of the resource. Experiment 2 looked at automatic classification of subjects, specifically how a document's keywords mapped to ScOT terms and to edna categories. Experiment 3 extended this investigation of automated subject analysis mapped to ScOT terms, but this time using a set of teacher-developed documents and presentations that had been tagged by a teacher, but had not been catalogued by an information professional.

Experiment 1: key phrase extraction

Problem experiment 1

Suggestions for resources to be considered for inclusion in edna may come from users via an email or a web-based suggestion form, or from subscription alerting services delivered via email or RSS feed. There is considerable potential for duplication of effort in handling suggested resources between staff in the edna metadata team, and if an automated system could confidently predict the intended audience of an incoming resource it could be directed to the most appropriate cataloguer for that education sector. If at the same time the subject of a new resource was identified automatically then the resource could be sent directly to the subject matter expert for that topic.

Could artificial intelligence techniques inform a system whereby resources identified for edna could be routed to the most appropriate edna metadata creator? Could keyword extraction determine whether key phrases from the text could be used to predict the subject (dc.subject), audience (edna-audience) and education sector (edna-sector) of a resource?

Method experiment 1

In order to inform the design of an automated process of key phrase extraction, it was necessary to collect data about how experts perform this task. To this end, edna staff used a web-based collection tool developed by Flinders University to review resources from the edna repository, attempting to select those particular snippets of text in a resource that best justified the allocation of subject terms. In a particular data collection session, a document from the existing edna repository would be chosen at random, and displayed along with previously captured metadata about subject terms, edna category and appropriate audience level. The information expert would then refer to the original text of the resource, and snip short textual excerpts from the original text to substantiate the metadata decisions made.

To date, some 300 documents have been examined in this way. This database of expert judgment information can potentially provide a wealth of information about how expert readers look for the most crucial information in the text of a resource for the purpose of classification. It may be possible to exploit these data to train an automatic computer algorithm to locate these key phrases automatically in an unseen resource, a necessary first step before performing text classification into various topic categories.

Results experiment 1

The small sample size that was possible from the human resources available for a proof of concept level project meant that meaningful results from this experiment could not be realised. Ongoing research in this way would be dependent on increased funding and the Flinders University team have been working to identify sources for funding to continue this. If resources become available, it is intended to increase the sample size and to try to devise techniques to automatically identify phrases thus 'training' the system. Developing a method of incorporating the snipping of key phrases into existing workflow for new resources, rather than the current tool which requires separate retrospective analysis, could facilitate the collection process.

The information professionals involved in locating keywords and phrases in original documents to support metadata decisions that had been made at the time of cataloguing became aware very quickly that while subject classification was relatively easy to support from key phrases in online documents, finding phrases that identified a specific audience group or user level was much more difficult. The words student and teacher appeared frequently together on education websites for both audience types and neither was a reliable indicator of intended user. When looking for indicators of user level on a resource, it was frequently non-text-based elements that researchers used such as the number and style of images on the site, or the size of the font. Often the most useful text-based clues were not on the home page or index page of the resource, but in a lower level page such as the About page. It would be interesting to test these preliminary instincts against data from future processing of key phrases.

Experiment 2: predict edna category from subject term using a semantic network approach

Problem experiment 2

Subject analysis is arguably the most important task in the cataloguing of education resources as it populates the metadata fields that are most useful to searchers. Subject analysis is the most time consuming task in the metadata creation workflow, and also requires the most professional expertise. While objective metadata fields such as title, publisher, format and date are more straightforward and can be completed with minimal training and subject domain knowledge, effective subject analysis requires both competence in use of controlled vocabularies and knowledge of the subject matter being analysed.

Could automatic classification of subjects be realised; specifically could a system automatically map a document's keywords to the most appropriate Schools Online Thesaurus (ScOT) terms and edna category?

Method experiment 2

During 2009, we conducted an experiment towards prototyping a topic classifier to associate subject terms provided for a sample of documents in the edna repository with the edna categories to which these

documents were assigned. While the "end-to-end" goal remained the implementation of a technique to extract key phrases from the text of a document and map the key phrases to controlled vocabulary terms, this experiment focussed on a related task that was more limited in scope but allowed us to hone the computational techniques involved in semantically-based text-to-text mappings in general. A set of edna school sector curriculum resources was used in this experiment, and as these resources had already been assigned ScOT and edna categories it was possible to compare the results of machine classification against how humans had classified the resources, and thus determine relative success rates.

The prototype classifier employed a *WordNet*-based semantic mapping scheme to map subject terms to categories, and two variants were developed. One variant employed only a standard mapping from subjects to edna categories using *WordNet*. The other variant also incorporated broad term expansion of the original subject term, i.e. first expanding the original term to a list of all terms that are synonymous with or semantically linked to the original term according to WordNet, and then searching for edna categories that were semantically related to this broader set of terms. This allowed for a wider range of *WordNet* relationships to be brought into play in order to select the edna category that was closest to the set of terms in *WordNet*. The calculation of semantic distance was done by means of a measure developed in the Flinders University Artificial Intelligence Laboratory to determine semantic distance in *WordNet* (Yang & Powers 2005, 2006).

Results experiment 2

The three text classification approaches were implemented and compared in terms of their accuracy in classifying unseen resources. It was found that attempting to predict the edna category from the subject term using semantic relatedness yielded very promising results, with a 60% success rate in mapping a subject term to a category (see Table 1), and 35% accuracy for mapping a category from the title of the resource alone (suggesting that the expert-provided subject term is a more reliable guide to the topic of the resource). Interestingly, adding broad term expansion reduced, rather than increased, accuracy. One reason seemed to be that, in cases where there simply was no matching edna category (which would have been the answer returned by the method without expansion), term expansion increased the opportunity for an incorrect mapping by including several terms that were semantically related to, but not synonymous with, the original term. Another reason was that the meanings of the original terms became 'diluted' by an expansion process that generated quite similar sets of related terms for semantically diverse words.

Table 1. Proportion of hits in the 10 top-ranked options when attempting to predict the edna category from various document features.

Subject without broad term expansion	60%
Subject with broad term expansion	45%
Title	35%

While these results are promising, their level of success is expected to improve with more sophisticated techniques and with the use of a richer underlying knowledge database, for example *Wikipedia* rather than *WordNet*.

Experiment 3 – mapping from keywords to ScOT (ST) words

Problem experiment 3

A stated goal of edna collection development at this time was to increase the proportion of teacher-created resources, for example lesson plans, online curriculum modules, presentation files and blog posts. This raised two issues for cataloguers:

1. the increased diversity of resource types beyond the websites and research literature that edna traditionally catalogued was a challenge, and

2. while cataloguers appreciated the value of harvesting of subject keywords through user tags, those who tagged these resources were not necessarily using ScOT controlled vocabulary terms, thus increasing inconsistency.

Could we extend the investigation of automated subject analysis mapped to ScOT terms, using a set of teacher-developed documents and presentations that have been tagged by a teacher, and not catalogued by information professionals?

Method experiment 3

The third experiment in this sequence of research attempted to use computer techniques to automatically annotate documents with categorical subject terms taken from a controlled vocabulary, namely the Schools Online Thesaurus (ScOT). As the document database, we made use of a set of resources provided by an edna partner organisation, the Victoria Information Technology Teachers' Association (VITTA). This data set was a sample of documents from the VITTA website and included conference handouts, presentations and lesson plans. These resources were tagged by a teacher with information technology subject expertise rather than by information specialists.

While the full-text documents were available for the research, the format of these documents being largely slide presentations or handouts which consisted of highly condensed textual information meant that automatic keyword extraction was not possible. There were insufficiently large differences in the frequencies with which different words occurred to allow us to easily identify the most important keywords. Hence, the research effort concentrated on the process of automatically mapping from the teacher-provided keywords to the controlled ScOT vocabulary. Three different approaches were examined and are described below. In order to evaluate the accuracy of each approach, the recall value, expressing the percentage of keywords that obtained a ScOT match under this approach, was calculated.

Results experiment 3

Full Mapping

The first point to be considered in effecting this mapping is that it is quite likely that many of the keywords provided by subject experts are in fact ScOT keywords. In order to derive a baseline level of accuracy for comparison with the more sophisticated semantic approaches, the recall value can be worked out for an approach that simply maps expert-supplied keywords to ScOT terms. In fact, 35.9% of all keywords in the sample matched ScOT terms. Example: "information literacy" → "Information Literacy".

Partial Mapping

A less strict approach, incorporating a modest amount of semantic information, is to create a mapping from a (single- or multiword) key phrase to any (single- or multiword) ScOT term that (i) contains any of the words in the phrase, or (ii) contains a synonym for any word in the phrase. Synonyms were obtained by consulting the *WordNet* thesaurus. With Partial Mapping, it was possible to achieve 43.1% recall. Example: "outcomes" → "Learning Outcomes".

Semantic Mapping

In contrast to the former two fairly simple approaches, Semantic Mapping attempts to connect keywords with ScOT terms based on deeper semantic links. Semantic Mapping makes use of the Full Mapping if it exists, but if it does not, a keyword is mapped onto the most closely semantically related term in ScOT. As there are both keywords and ScOT terms that do not occur in *WordNet*, this technique cannot be expected to map all keywords to ScOT terms. Nevertheless, Semantic Mapping achieves 22.8% recall. This number is not as high as was achieved with the Full or Partial Mapping approaches, but provides the ability to go well beyond what can be achieved with these simplistic textual-match approaches alone. Semantic Mapping potentially identifies ScOT terms that are not in any way textually related to their keywords, and so provides the ability to augment the results of an initial Full or Partial Mapping phase with a deeper, semantically-

 Diversity Challenge Resilience: School Libraries in Action Proceedings of the 12th Biennial School Library Association of Queensland, the 39th International Association of School Librarianship Annual Conference incorporating the 14th International Forum on Research in School Librarianship, Brisbane QLD Australia, 27 September – 1 October 2010.

based mapping phase. Examples: "worms" → "Computer Programs"; "LAN" → "Computer Networks"; "article" → "Non-fiction."

Table 2. Recall values (percentage of ScOT terms correctly predicted from keywords) for three different keyword-ScOT term mapping approaches.

Mapping Approach	Recall
Full Mapping	35.9%
Partial Mapping	43.1%
Semantic Mapping	22.8%

Some additional insights were arrived at through the mapping exercise. It was noted that the names of several specific information technology platforms and programming languages that were seen as important keywords to the subject expert IT teacher, were not included in ScOT, for instance ".NET", "laptop", "phishing", "iPod", "Excel", and "podcast".

Also, several mappings that were achieved between keywords and ScOT terms were based on incorrect semantic associations, as a result of linking an ambiguous keyword with a word related to an unintended meaning. Example: "notes" → "Pitch (Music)". On checking the context of this keyword it became clear that this referred to a resource about note-taking for examination purposes, not to a resource about music.

In other cases the mapping techniques linked words together that had a stem in common but were in fact quite different in meaning, for example "practice exams" was linked to "Practical Examinations" which does not refer to exactly the same concept. This seems to have occurred because words were first reduced to their stems, i.e. the essential core form of a word, by removing any of the common affixes and suffixes in English ("-ing", "-er", "un-"). In the case above, the words "practice" and "practical" were both reduced to the stem "pract-", thereby causing them to be incorrectly treated as the same concept.

It is envisaged that a more complex handling of semantic relatedness may be able to extend the foundation established by this research, by

1. taking into account the inherent ambiguity of words,
2. going beyond the use of word stems by incorporating the fine-grained semantic information provided by affixes and suffixes, and
3. making use of a richer source of real-world knowledge such as Wikipedia.

Discussion

There were several limitations which impacted on this investigation, including the nature of funding which although a common issue for proof of concepts, in this case meant there was more emphasis on methodology than on in-depth analysis of results. The first priority was to test whether these investigations could realistically return useful results in the event that funding for a full research project could be obtained. We learned that experiment 1 was very demanding in terms of cataloguer time and expertise, but invaluable in raising awareness of the metadata creation workflow and the fact that audience and user level metadata decisions were significantly more complex than subject metadata. This experiment would benefit from re-design to maximise efficiency of research time and effort. It was also unfortunate that we were not in a position to directly compare the results of experiments 2 and 3. This would have required analysis of the same set of documents, whereas we conducted these investigations with two distinct datasets. It might be tempting to infer that because one of the automated subject classification techniques based on full text documents in experiment 2 returned a much higher accuracy (60%) than the mapping between subject-specialist keywords and ScOT (43.1%), therefore the keyword tagging of resources is not warranted. However, there were so many differences between the datasets used in these experiments that it is not possible to compare them, including the fact that the information technology discipline which was the focus of experiment 3, is an area to which the ScOT thesaurus has not to this stage given a high priority.

A key finding of the project is that automated classification based on artificial intelligence is useful as a means of supplementing and assisting human classification, but is not at this stage a replacement for human classification. In any software, workflow or process developed using automated classification, there is still a requirement to build in the ability for information management staff to be able to moderate any suggestions automatically made by the software, particularly to deal with issues of word stemming, disambiguation of words from different disciplines and highly specific or recent terminology.

The results from experiments 2 and 3 are in line with previous work in the text classification literature (Budanitsky & Hirst, 2006; Pedersen et al., 2004) showing that semantic information from a structured ontology such as *WordNet* can allow classifiers to go beyond simple textual correspondences between texts and terms, to find correspondences at the level of underlying meaning. This project also confirmed the usefulness of this kind of semantic mapping in the specific domain of education-related text resources.

The research findings could inform the development of new collection and cataloguing workflows for those working with digital content. Specifically for the work of digital repositories such as edna, this project could assist by

1. analysing the text of new resources in order to automatically classify them into edna topics,
2. suggesting controlled vocabulary subject terms for new resources, including user tagged resources and
3. potentially automatically searching for new resources for inclusion in edna based on identified gaps in the collection.

The results will be of interest to those involved in the building and management of repositories or thesauri, dealing with the current challenges presented by the need to identify, evaluate and describe resources of high quality and relevance to users, from the vast numbers of items available.

Further development of artificial intelligence tools should be undertaken to automate identification, keyword extraction and related works, thus freeing the human cataloguers to do the tasks they do best: high level evaluation, analysis, interpretation and educational description of resources.

Conclusion

The project aimed to determine to what extent artificial intelligence techniques can be employed to improve the efficiency of digital collection building and to provide greater consistency of learning resource metadata. As a flip side of this it addresses the question of which elements of collection management are best performed by human specialists. The results show in terms of the research question about which metadata creation decisions artificial intelligence systems could be trained to implement, that we have had some success with subject categorisation on our small sample (see Table 1). In terms of keyword extraction to inform audience or sector elements, we have insufficient data and require further human input to proceed in investigation of this area.

This research sought to improve and expedite the process of collecting learning resources and describing them for inclusion in a digital repository. The project also sought to contribute to scientific knowledge of word meaning-based techniques of automated text classification. An associated benefit of the research is its examination of metadata creation behaviours and review of cognitive behavioural activities which will prove very useful in training of staff involved in metadata creation. The research findings could inform the development of new collection and cataloguing workflows for those working with digital content.

Three key learnings:

1. Artificial intelligence systems showed some success in subject categorisation of text-based digital learning resources
2. Key phase extraction to support categorisation of audience and user-level is more difficult than subject categorisation
3. Automated classification based on artificial intelligence may be useful as a means of supplementing and assisting human classification, but is not at this stage a replacement for human classification

References

Budanitsky, A. & Hirst, G. (2006). Evaluating WordNet-based measures of lexical semantic relatedness. *Computational Linguistics*, 32(1), 13-47.

Cardinaels, K., Meire, M., & Duval, E. (2005). Automating metadata generation: the simple indexing interface. In *Proceedings of the 14th International Conference on World Wide Web WWW '05,* pp. 548-556.

Cotton, T. (2007). Defining proof of concepts. *Tom Cotton's 8109 blog*. Education.au. Retrieved April 15, 2010, from http://blogs.educationau.edu.au/tcotton/2007/10/08/defining-proof-of-concepts

Dublin Core Metadata Initiative (1998). *Dublin Core metadata element set, version 1.0: Reference description*. Retrieved April 15, 2010, from http://www.dublincore.org/documents/1998/09/dces

Education Network Australia. (2001). *edna metadata standard v.1.1*. Retrieved, April 15, 2010, from http://www.edna.edu.au/edna/go/resources/metadata/pid/261

Education Network Australia. (2006). *edna resources metadata application profile 1.0*. Retrieved April 15, 2010, from http://www.edna.edu.au/edna/webdav/site/myjahiasite/shared/edna_resources_metadata_1.0.pdf

Education Network Australia. (2007). *edna collection policy*. Retrieved April 15, 2010, from http://www.edna.edu.au/edna/go/about/policies/collection

Education Services Australia. (2010). *Schools Online Thesaurus (ScOT)*. Retrieved April 15, 2010, from http://scot.curriculum.edu.au

Education.au. (2007). *Events metadata application profile v 1.3*. Retrieved April 15, 2010, from http://www.edna.edu.au/edna/webdav/site/myjahiasite/shared/edna_events_metadata_1.3.pdf

Farrell, J. & Lausen, H. (2007). *Semantic annotation for WSDL and XML schema,* W3C Semantic Annotations for Web Service Description Language Working Group, Retrieved April 15, 2010, from http://www.w3.org/TR/sawsdl

Fellbaum, C. (1998). *WordNet: an electronic lexical database*, Cambridge, MA: The MIT Press.

Greenberg, J., Spurgin, K. & Crystal, A. (2006). Functionalities for automatic metadata generation applications: a survey of metadata experts' opinions. *International Journal of Metadata, Semantics and Ontologies*, 1(1) 3-20.

Hayman, S. & Lothian, N. (2009). Towards linked education data: metadata extraction projects for Education Network Australia (edna). *Proceedings of the International Conference on Dublin Core and Metadata Applications,* Retrieved April 15, 2010, from http://dcpapers.dublincore.org/ojs/pubs/article/view/978/958

Lothian, N. (2007). Folksonomies, people and manufactured serendipity. *Education.au blog,* Retrieved April 15, 2010, from http://blogs.educationau.edu.au/nlothian/2007/07/06/folksonomies-people-and-manufactured-serendipity

Ochoa, X. & Duval, E. (2007). Towards automatic evaluation of metadata quality in digital repositories. *Ariadne,* Retrieved April 15, 2010, from http://ariadne.cti.espol.edu.ec/M4M/files/TowardsAutomaticQuality.pdf

Paynter, G. (2005). Developing practical automatic metadata assignment. *JCDL '05: Proceedings of the 5th ACM/IEEE-CS Joint Conference on Digital Libraries,* pp. 291-300.

Pedersen, T., Patwardhan, S. & Michelizzi, J. (2004). Wordnet: Similarity - measuring the relatedness of concepts. *Proceedings of the Nineteenth National Conference on Artificial Intelligence (AAAI-04)*, pp. 1024–1025.

Pfitzner, D., Treharne, K. & Powers, D. (2008). User keyword preference: the Nwords and Rwords experiments. *International Journal of Internet Protocol Technology*, 9:149-158.

Ponzetto, S. & Strube, M. (2007). Knowledge derived from Wikipedia for computing semantic relatedness. *Journal of Artificial Intelligence Research*, 30(1), 181-212, Retrieved April 15, 2010, from http://www.aaai.org/Papers/JAIR/Vol30/JAIR-3005.pdf

Princeton University. (2010). *WordNet*. Retrieved April 15, 2010, from http://wordnet.princeton.edu

Ramos, J. (2003). Using TF-IDF to determine word relevance in document queries. *The First instructional Conference on Machine Learning (iCML-2003)*, Retrieved April 15, 2010, from http://www.cs.rutgers.edu/~mlittman/courses/ml03/iCML03/papers/ramos.pdf

Rowlands, D. (2004). *Real world metadata management for resource discovery: proof of concept across education and library sectors in Tasmania*. University of Tasmania ePrints. Retrieved April 15, 2010, from http://eprints.utas.edu.au/15/1/Real_World_Metadata.pdf

Sebastiani, F. (2002). Machine learning in automated text categorization. *ACM Computing Surveys (CSUR)*, 34(1), 1-47.

Turdakov, D. & Velikhov, P. (2008). Semantic relatedness metric for Wikipedia concepts based on link analysis and its application to word sense disambiguation. In S. D. Kuznetsov, P. Pleshachkov, B. Novikov, and D. Shaporenkov *Proceedings of SYRCoDIS,* Volume 355 of CEUR Workshop*,* Retrieved April 15, 2010, from http://sunsite.informatik.rwth-aachen.de/Publications/CEUR-WS/Vol-355/turdakov.pdf

Wikimedia Foundation. (2010). *Wikipedia.* Retrieved April 15, 2010, from http://en.wikipedia.org

Yang, D. & Powers, D. (2005). Measuring semantic similarity in the taxonomy of WordNet. *ACSC '05: Proceedings of the Twenty-eighth Australasian conference on Computer Science,* Darlinghurst, Australia, pp. 315-322.

Yang, D. & Powers, D. (2006). Verb similarity on the taxonomy of WordNet. *The Third International WordNet Conference (GWC-06),* Jeju Island, Korea, pp. 121-129.

Statement of Originality

This statement certifies that the paper above is based upon original research undertaken by the authors and that the paper was conceived and written by the authors alone and has not been published elsewhere. All information and ideas from others is referenced.