SGAI

AI
2016

Max Bramer
Miltos Petridis *Editors*

Research and Development in Intelligent Systems XXXIII

Incorporating Applications and Innovations in Intelligent Systems XXIV

Proceedings of AI-2016,
The Thirty-Sixth SGAI International
Conference on Innovative Techniques
and Applications of Artificial
Intelligence

Springer

Research and Development in Intelligent Systems XXXIII

Incorporating Applications and Innovations in Intelligent Systems XXIV

Max Bramer · Miltos Petridis
Editors

Research and Development in Intelligent Systems XXXIII

Incorporating Applications and Innovations in Intelligent Systems XXIV

Proceedings of AI-2016, The Thirty-Sixth SGAI International Conference on Innovative Techniques and Applications of Artificial Intelligence

Springer

Editors
Max Bramer
School of Computing
University of Portsmouth
Portsmouth
UK

Miltos Petridis
School of Computing, Engineering
and Mathematics
University of Brighton
Brighton
UK

ISBN 978-3-319-47174-7 ISBN 978-3-319-47175-4 (eBook)
DOI 10.1007/978-3-319-47175-4

Library of Congress Control Number: 2016954594

Printed on acid-free paper

This Springer imprint is published by Springer Nature
The registered company is Springer International Publishing AG
The registered company address is: Gewerbestrasse 11, 6330 Cham, Switzerland

Programme Chairs' Introduction

This volume comprises the refereed papers presented at AI-2016, the Thirty-sixth SGAI International Conference on Innovative Techniques and Applications of Artificial Intelligence, held in Cambridge in December 2016 in both the technical and the application streams. The conference was organised by SGAI, the British Computer Society Specialist Group on Artificial Intelligence.

The technical papers included present new and innovative developments in the field, divided into sections on Knowledge Discovery and Data Mining, Sentiment Analysis and Recommendation, Machine Learning, AI Techniques, and Natural Language Processing. This year's Donald Michie Memorial Award for the best-refereed technical paper was won by a paper entitled "Harnessing Background Knowledge for E-learning Recommendation" by B. Mbipom, S. Craw and S. Massie (Robert Gordon University, Aberdeen, UK).

The application papers included present innovative applications of AI techniques in a number of subject domains. This year, the papers are divided into sections on legal liability, medicine and finance, telecoms and e-Learning, and genetic algorithms in action. This year's Rob Milne Memorial Award for the best-refereed application paper was won by a paper entitled "A Genetic Algorithm Based Approach for the Simultaneous Optimisation of Workforce Skill Sets and Team Allocation" by A.J. Starkey and H. Hagras (University of Essex, UK), S. Shakya and G. Owusu (British Telecom, UK).

The volume also includes the text of short papers presented as posters at the conference.

On behalf of the conference organising committee, we would like to thank all those who contributed to the organisation of this year's programme, in particular the programme committee members, the executive programme committees and our administrators Mandy Bauer and Bryony Bramer.

Max Bramer, Technical Programme Chair, AI-2016
Miltos Petridis, Application Programme Chair, AI-2016

Acknowledgements/Committees

AI-2016 Conference Committee

Prof. Max Bramer, University of Portsmouth (Conference Chair)
Prof. Max Bramer, University of Portsmouth (Technical Programme Chair)
Prof. Miltos Petridis, University of Brighton (Application Programme Chair)
Dr. Jixin Ma, University of Greenwich (Deputy Application Programme Chair)
Prof. Adrian Hopgood, University of Liege, Belgium (Workshop Organiser)
Rosemary Gilligan (Treasurer)
Dr. Nirmalie Wiratunga, Robert Gordon University, Aberdeen (Poster Session Organiser)
Andrew Lea, Primary Key Associates Ltd. (AI Open Mic and Panel Session Organiser)
Dr. Frederic Stahl, University of Reading (Publicity Organiser)
Dr. Giovanna Martinez, Nottingham Trent University and Christo Fogelberg Palantir Technologies (FAIRS 2016)
Prof. Miltos Petridis, University of Brighton and Prof. Thomas Roth-Berghofer University of West London (UK CBR Organisers)
Mandy Bauer, BCS (Conference Administrator)
Bryony Bramer, (Paper Administrator)

Technical Executive Programme Committee

Prof. Max Bramer, University of Portsmouth (Chair)
Prof. Frans Coenen, University of Liverpool
Dr. John Kingston, University of Brighton
Prof. Dan Neagu, University of Bradford
Prof. Thomas Roth-Berghofer, University of West London
Dr. Nirmalie Wiratunga, Robert Gordon University, Aberdeen

Applications Executive Programme Committee

Prof. Miltos Petridis, University of Brighton (Chair)
Mr. Richard Ellis, Helyx SIS Ltd.
Ms. Rosemary Gilligan, University of Hertfordshire
Dr. Jixin Ma, University of Greenwich (Vice-Chair)
Dr. Richard Wheeler, University of Edinburgh

Technical Programme Committee

Andreas Albrecht (Middlesex University)
Abdallah Arioua (IATE INRA France)
Raed Batbooti (University of Swansea UK (PhD Student), University of Basra (Lecturer))
Lluís Belanche (Universitat Politecnica de Catalunya, Barcelona, Catalonia, Spain)
Yaxin Bi (Ulster University, UK)
Mirko Boettcher (University of Magdeburg; Germany)
Max Bramer (University of Portsmouth)
Krysia Broda (Imperial College; University of London)
Ken Brown (University College Cork)
Charlene Cassar (De Montfort University UK)
Frans Coenen (University of Liverpool)
Ireneusz Czarnowski (Gdynia Maritime University; Poland)
Nicolas Durand (Aix-Marseille University)
Frank Eichinger (CTS EVENTIM AG & Co. KGaA, Hamburg, Germany)
Mohamed Gaber (Robert Gordon University, Aberdeen, UK)
Hossein Ghodrati Noushahr (De Montfort University, Leicester, UK)
Wael Hamdan (MIMOS Berhad., Kuala Lumpur, Malaysia)
Peter Hampton (Ulster University, UK)
Nadim Haque (Capgemini)
Chris Headleand (University of Lincoln, UK)
Arjen Hommersom (Open University, The Netherlands)
Adrian Hopgood (University of Liège, Belgium)
John Kingston (University of Brighton)
Carmen Klaussner (Trinity College Dublin Ireland)
Ivan Koychev (University of Sofia)
Thien Le (University of Reading)
Nicole Lee (University of Hong Kong)
Anne Liret (British Telecom France)
Fernando Lopes (LNEG-National Research Institute; Portugal)
Stephen Matthews (Newcastle University)
Silja Meyer-Nieberg (Universitat der Bundeswehr Munchen Germany)

Roberto Micalizio (Universita' di Torino)
Daniel Neagu (University of Bradford)
Lars Nolle (Jade University of Applied Sciences; Germany)
Joanna Isabelle Olszewska (University of Gloucestershire UK)
Dan O'Leary (University of Southern California)
Juan Jose Rodriguez (University of Burgos)
Thomas Roth-Berghofer (University of West London)
Fernando Saenz-Perez (Universidad Complutense de Madrid)
Miguel A. Salido (Universidad Politecnica de Valencia)
Rainer Schmidt (University Medicine of Rostock; Germany)
Frederic Stahl (University of Reading)
Simon Thompson (BT Innovate)
Jon Timmis (University of York)
M.R.C. van Dongen (University College Cork)
Martin Wheatman (Yagadi Ltd.)
Graham Winstanley (University of Brighton)
Nirmalie Wiratunga (Robert Gordon University)

Application Programme Committee

Hatem Ahriz (Robert Gordon University)
Tony Allen (Nottingham Trent University)
Ines Arana (Robert Gordon University)
Mercedes Arguello Casteleiro (University of Manchester)
Ken Brown (University College Cork)
Sarah Jane Delany (Dublin Institute of Technology)
Richard Ellis (Helyx SIS Ltd.)
Roger Evans (University of Brighton)
Andrew Fish (University of Brighton)
Rosemary Gilligan (University of Hertfordshire)
John Gordon (AKRI Ltd.)
Chris Hinde (Loughborough University)
Adrian Hopgood (University of Liege, Belgium)
Stelios Kapetanakis (University of Brghton)
Alice Kerly
Jixin Ma (University of Greenwich)
Lars Nolle (Jade University of Applied Sciences)
Miltos Petridis (University of Brighton)
Miguel A. Salido (Universidad Politecnica de Valencia)
Roger Tait (University of Cambridge)
Richard Wheeler (Edinburgh Scientific)

Contents

Research and Development in Intelligent Systems XXXIII

Best Technical Paper

Knowledge Discovery and Data Mining

Sentiment Analysis and Recommendation

Machine Learning

AI Techniques

Natural Language Processing

Short Papers

Applications and Innovations in Intelligent Systems XXIV

Best Application Paper

Legal Liability, Medicine and Finance

Telecoms and E-Learning

Genetic Algorithms in Action

Short Papers

Research and Development in Intelligent Systems XXXIII

Best Technical Paper

Harnessing Background Knowledge for E-Learning Recommendation

Blessing Mbipom, Susan Craw and Stewart Massie

Abstract The growing availability of good quality, learning-focused content on the Web makes it an excellent source of resources for e-learning systems. However, learners can find it hard to retrieve material well-aligned with their learning goals because of the difficulty in assembling effective keyword searches due to both an inherent lack of domain knowledge, and the unfamiliar vocabulary often employed by domain experts. We take a step towards bridging this semantic gap by introducing a novel method that automatically creates custom background knowledge in the form of a set of rich concepts related to the selected learning domain. Further, we develop a hybrid approach that allows the background knowledge to influence retrieval in the recommendation of new learning materials by leveraging the vocabulary associated with our discovered concepts in the representation process. We evaluate the effectiveness of our approach on a dataset of Machine Learning and Data Mining papers and show it to outperform the benchmark methods.

Keywords Knowledge Discovery · Recommender Systems · eLearning Systems · Text Mining

1 Introduction

There is currently a large amount of e-learning resources available to learners on the Web. However, learners have insufficient knowledge of the learning domain, and are not able to craft good queries to convey what they wish to learn. So, learners are

B. Mbipom · S. Craw · S. Massie (✉)
School of Computing Science and Digital Media, Robert Gordon University,
Aberdeen, UK
e-mail: s.massie@rgu.ac.uk

B. Mbipom
e-mail: b.e.mbipom@rgu.ac.uk

S. Craw
e-mail: s.craw@rgu.ac.uk

M. Bramer and M. Petridis (eds.), *Research and Development in Intelligent Systems XXXIII*, DOI 10.1007/978-3-319-47175-4_1

often discouraged by the time spent in finding and assembling relevant resources to meet their learning goals [5]. E-learning recommendation offers a possible solution.

E-learning recommendation typically involves a learner query, as an input; a collection of learning resources from which to make recommendations; and selected resources recommended to the learner, as an output. Recommendation differs from an information retrieval task because with the latter, the user requires some understanding of the domain in order to ask and receive useful results, but in e-learning, learners do not know enough about the domain. Furthermore, the e-learning resources are often unstructured text, and so are not easily indexed for retrieval [11]. This challenge highlights the need to develop suitable representations for learning resources in order to facilitate their retrieval.

We propose the creation of background knowledge that can be exploited for problem-solving. In building our method, we leverage the knowledge of instructors contained in eBooks as a guide to identify the important domain topics. This knowledge is enriched with information from an encyclopedia source and the output is used to build our background knowledge. DeepQA applies a similar approach to reason on unstructured medical reports in order to improve diagnosis [9]. We demonstrate the techniques in Machine Learning and Data Mining, however the techniques we describe can be applied to other learning domains.

In this paper, we build background knowledge that can be employed in e-learning environments for creating representations that capture the important concepts within learning resources in order to support the recommendation of resources. Our method can also be employed for query expansion and refinement. This would allow learners' queries to be represented using the vocabulary of the domain with the aim of improving retrieval. Alternatively, our approach can enable learners to browse available resources through a guided view of the learning domain.

We make two contributions: firstly, the creation of background knowledge for an e-learning domain. We describe how we take advantage of the knowledge of experts contained in eBooks to build a knowledge-rich representation that is used to enhance recommendation. Secondly, we present a method of harnessing background knowledge to augment the representation of learning resources in order to improve the recommendation of resources. Our results confirm that incorporating background knowledge into the representation improves e-learning recommendation.

This paper is organised as follows: Sect. 2 presents related methods used for representing text; Sect. 3 describes how we exploit information sources to build our background knowledge; Sect. 4 discusses our methods in harnessing a knowledge-rich representation to influence e-learning recommendation; and Sect. 5 presents our evaluation. We conclude in Sect. 6 with insights to further ways of exploiting our background knowledge.

2 Related Work

Finding relevant resources to recommend to learners is a challenge because the resources are often unstructured text, and so are not appropriately indexed to support the effective retrieval of relevant materials. Developing suitable representations to improve the retrieval of resources is a challenging task in e-learning environments [8], because the resources do not have a pre-defined set of features by which they can be indexed. So, e-learning recommendation requires a representation that captures the domain-specific vocabulary contained in learning resources. Two broad approaches are often used to address the challenge of text representation: corpus-based methods such as topic models [6], and structured representations such as those that take advantage of ontologies [4].

Corpus-based methods involve the use of statistical models to identify topics from a corpus. The identified topics are often keywords [2] or phrases [7, 18]. Coenen et al. showed that using a combination of keywords and phrases was better than using only keywords [7]. Topics can be extracted from different text sources such as learning resources [20], metadata [3], and Wikipedia [14]. One drawback of the corpus-based approach is that, it is dependent on the document collection used, so the topics produced may not be representative of the domain. A good coverage of relevant topics is required when generating topics for an e-learning domain, in order to offer recommendations that meet learners' queries which can be varied.

Structured representations capture the relationships between important concepts in a domain. This often entails using an existing ontology [11, 15], or creating a new one [12]. Although ontologies are designed to have a good coverage of their domains, the output is still dependent on the view of its builders, and because of handcrafting, existing ontologies cannot easily be adapted to new domains. E-learning is dynamic because new resources are becoming available regularly, and so using fixed ontologies limits the potential to incorporate new content.

A suitable representation for e-learning resources should have a good coverage of relevant topics from the domain. So, the approach in this paper draws insight from the corpus-based methods and structured representations. We leverage on a structured corpus of teaching materials as a guide for identifying important topics within an e-learning domain. These topics are a combination of keywords and phrases as recommended in [7]. The identified topics are enriched with discovered text from Wikipedia, and this extends the coverage and richness of our representation.

3 Background Knowledge Representation

Background knowledge refers to information about a domain that is useful for general understanding and problem-solving [21]. We attempt to capture background knowledge as a set of domain concepts, each representing an important topic in the domain. For example, in a learning domain, such as Machine Learning, you would

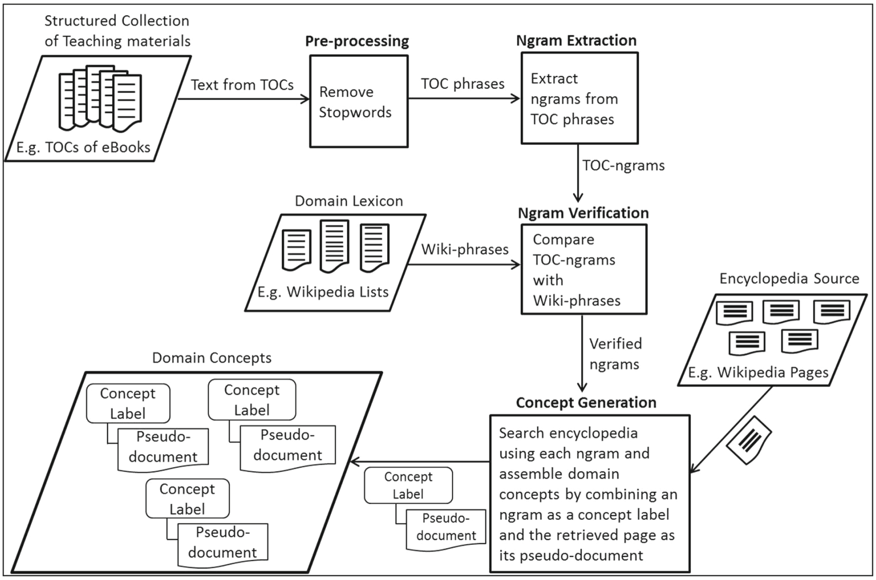


Fig. 1 An overview of the background knowledge creation process

find topics such as Classification, Clustering and Regression. Each of these topics would be represented by a concept, in the form of a concept label and a pseudo-document which describes the concept. The concepts can then be used to underpin the representation of e-learning resources.

The process involved in discovering our set of concepts is illustrated in Fig. 1. Domain knowledge sources are required as an input to the process, and we use a structured collection of teaching materials and an encyclopedia source. We automatically extract ngrams from our structured collection to provide a set of potential concept labels, and then we use a domain lexicon to validate the extracted ngrams in order to ensure that the ngrams are also being used in another information source. The encyclopedia provides candidate pages that become the concept label and discovered text for the ngrams. The output from this process is a set of concepts, each comprising a label and an associated pseudo-document. The knowledge extraction process is discussed in more detail in the following sections.

3.1 Knowledge Sources

Two knowledge sources are used as initial inputs for discovering concept labels. A structured collection of teaching materials provides a source for extracting important topics identified by teaching experts in the domain, while a domain lexicon provides a broader but more detailed coverage of the relevant topics in the domain. The lexicon is

Table 1 Summary of eBooks used

Book title and author	Cites
Machine learning; Mitchell	264
Introduction to machine learning; Alpaydin	2621
Machine learning a probabilistic perspective; Murphy	1059
Introduction to machine learning; Kodratoff	159
Gaussian processes for machine learning; Rasmussen and Williams	5365
Introduction to machine learning; Smola and Vishwanathan	38
Machine learning, neural and statistical classification; Michie, Spiegelhalter, and Taylor	2899
Introduction to machine learning; Nilsson	155
A first encounter with machine learning; Welling	7
Bayesian reasoning and machine learning; Barber	271
Foundations of machine learning; Mohri, Rostamizadeh, and Talwalkar	197
Data mining-practical machine learning tools and techniques; Witten and Frank	27098
Data mining concepts models and techniques; Gorunescu	244
Web data mining; Liu	1596
An introduction to data mining; Larose	1371
Data mining concepts and techniques; Han and Kamber	22856
Introduction to data mining; Tan, Steinbach, and Kumar	6887
Principles of data mining; Bramer	402
Introduction to data mining for the life sciences; Sullivan	15
Data mining concepts methods and applications; Yin, Kaku, Tang, and Zhu	23

used to verify that the concept labels identified from the teaching materials are directly relevant. Thereafter, an encyclopedia source, such as Wikipedia pages, is searched and provides the relevant text to form a pseudo-document for each verified concept label. The final output from this process is our set of concepts each comprising a concept label and an associated pseudo-document.

Our approach is demonstrated with learning resources from Machine Learning and Data Mining. We use eBooks as our collection of teaching materials; a summary of the books used is shown in Table 1. Two Google Scholar queries: "Introduction to data mining textbook" and "Introduction to machine learning textbook" guided the selection process, and 20 eBooks that meet all of the following 3 criteria were chosen. Firstly, the book should be about the domain. Secondly, there should be Google Scholar citations for the book. Thirdly, the book should be accessible. We use the Tables-of-Contents (TOCs) of the books as our structured knowledge source.

We use Wikipedia to create our domain lexicon because it contains articles for many learning domains [17], and the contributions of many people [19], so this provides the coverage we need in our lexicon. The lexicon is generated from 2 Wikipedia sources. First, the phrases in the *contents* and *overview* sections of the chosen domain are extracted to form a topic list. In addition, a list containing the titles of articles related to the domain is added to the topic list to assemble our lexicon. Overall, our domain lexicon consists of a set of 664 Wiki-phrases.

3.2 Generating Potential Domain Concept Labels

In the first stage of the process, the text from the TOCs is pre-processed. We remove characters such as punctuation, symbols, and numbers from the TOCs, so that only words are used for generating concept labels. After this, we remove 2 sets of stopwords. First, a standard English stopwords list,[1] which allows us to remove common words and still retain a good set of words for generating our concept labels. Our second stopwords are an additional set of words which we refer to as TOC-stopwords. It contains: structural words, such as *chapter* and *appendix*, which relate to the structure of the TOCs; roman numerals, such as *xxiv* and *xxxv*, which are used to indicate the sections in a TOC; and words, such as *introduction* and *conclusion*, which describe parts of a learning material and are generic across domains.

We do not use stemming because we found it harmful during pre-processing. When searching an encyclopedia source with the stemmed form of words, relevant results would not be returned. In addition, we intend to use the background knowledge for query refinement, so stemmed words would not be helpful.

The output from pre-processing is a set of TOC phrases. In the next stage, we apply ngram extraction to the TOC phrases to generate all 1–3 grams across the entire set of TOC phrases. The output from this process are TOC-ngrams containing a set of 2038 unigrams, 5405 bigrams and 6133 trigrams, which are used as the potential domain concept labels. Many irrelevant ngrams are generated from the TOCs because we have simply selected all 1–3 grams.

3.3 Verifying Concept Labels Using Domain Lexicon

The TOC-ngrams are first verified using a domain lexicon to confirm which of the ngrams are relevant for the domain. Our domain lexicon contains a set of 664 Wiki-phrases, each of which is pre-processed by removing non-alphanumeric characters. The 84 % of the Wiki-phrases that are 1–3 grams are used for verification. The comparison of TOC-ngrams with the domain lexicon identifies the potential domain concept labels that are actually being used to describe aspects of the chosen domain in Wikipedia. During verification, ngrams referring directly to the title of the domain, e.g. *machine learning* and *data mining*, are not included because our aim is to generate concept labels that describe the topics within the domain. In addition, we intend to build pseudo-documents describing the identified labels, and so using the title of

[1] http://snowball.tartarus.org/algorithms/english/stop.txt.

the domain would refer to the entire domain rather than specific topics. Overall, a set of 17 unigrams, 58 bigrams and 15 trigrams are verified as potential concept labels. Bigrams yield the highest number of ngrams, which indicates that bigrams are particularly useful for describing topics in this domain.

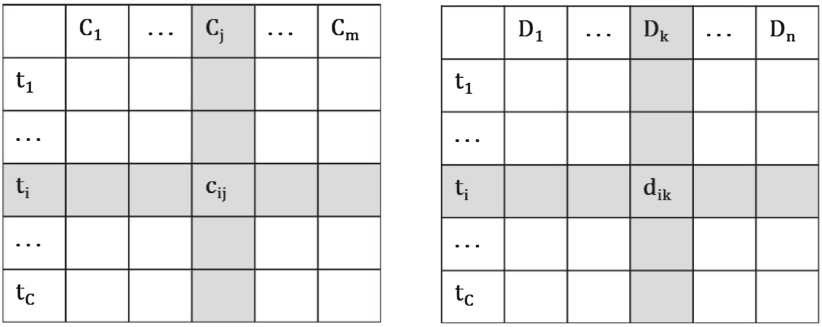
3.4 Domain Concept Generation

Our domain concepts are generated after a second verification step is applied to the ngrams returned from the previous stage. Each ngram is retained as a concept label if all of 3 criteria are met. Firstly, if a Wikipedia page describing the ngram exists. Secondly, if the text describing the ngram is not contained as part of the page describing another ngram. Thirdly, if the ngram is not a synonym of another ngram. For the third criteria, if two ngrams are synonyms, the ngram with the higher frequency is retained as a concept label while its synonym is retained as part of the extracted text. For example, 2 ngrams *cluster analysis* and *clustering* are regarded as synonyms in Wikipedia, so the text associated with them is the same. The label *clustering* is retained as the concept label because it occurs more frequently in the TOCs, and its synonym, *cluster analysis* is contained as part of the discovered text.

The concept labels are used to search Wikipedia pages in order to generate a domain concept. The search returns discovered text that forms a pseudo-document which includes the concept label. The concept label and pseudo-document pair make up a domain concept. Overall, 73 domain concepts are generated. Each pseudo-document is pre-processed using standard techniques such as removal of English stopwords and Porter stemming [13]. The terms from the pseudo-documents form the concept vocabulary that is now used to represent learning resources.

4 Representation Using Background Knowledge

Our background knowledge contains a rich representation of the learning domain and by harnessing this knowledge for representing learning resources, we expect to retrieve documents based on the domain concepts that they contain. The domain concepts are designed to be effective for e-learning, because they are assembled from the TOCs of teaching materials [1]. This section presents two approaches which have been developed by employing our background knowledge in the representation of learning resources.



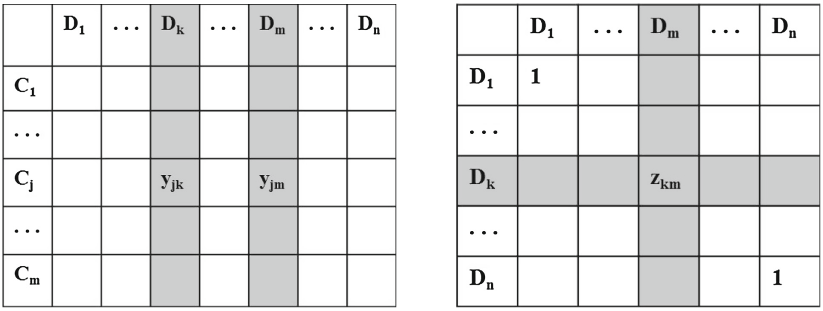
	C_1	…	C_j	…	C_m
t_1					
…					
t_i			c_{ij}		
…					
t_C					

(a) Term-concept matrix

	D_1	…	D_k	…	D_n
t_1					
…					
t_i			d_{ik}		
…					
t_C					

(b) Term-document matrix

Fig. 2 Term matrices for concepts and documents



	D_1	…	D_k	…	D_m	…	D_n
C_1							
…							
C_j			y_{jk}		y_{jm}		
…							
C_m							

(a) Concept-document matrix representation

	D_1	…	D_m	…	D_n
D_1	1				
…					
D_k			z_{km}		
…					
D_n					1

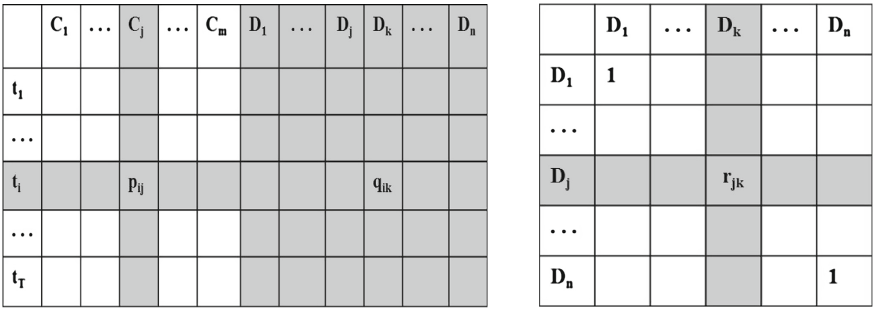
(b) Document-document similarity

Fig. 3 Document representation and similarity using the ConceptBased approach

4.1 The ConceptBased Approach

Representing documents with the concept vocabulary allows retrieval to focus on the concepts contained in the documents. Figures 2 and 3 illustrate the ConceptBased method. Firstly, in Fig. 2, the concept vocabulary, $t_1 \ldots t_c$, from the pseudo-documents of concepts, $C_1 \ldots C_m$, is used to create a term-concept matrix and a term-document matrix using TF-IDF weighting [16]. In Fig. 2a, c_{ij} is the TF-IDF of term t_i in concept C_j, while Fig. 2b shows d_{ik} which is the TF-IDF of t_i in D_k.

Next, documents D_1 to D_n are represented with respect to concepts by computing the cosine similarity of the term vectors for concepts and documents. The output is the concept-document matrix shown in Fig. 3a, where y_{jk} is the cosine similarity of the vertical shaded term vectors for C_j and D_k from Fig. 2a, b respectively. Finally, the document similarity is generated by computing the cosine similarity of concept-vectors for documents. Figure 3b shows z_{km}, which is the cosine similarity of the concept-vectors for D_k and D_m from Fig. 3a.



	C_1	…	C_j	…	C_m	D_1	…	D_j	D_k	…	D_n
t_1											
…											
t_i			p_{ij}						q_{ik}		
…											
t_T											

(a) Hybrid term-document matrix representation

	D_1	…	D_k	…	D_n
D_1	1				
…					
D_j			r_{jk}		
…					
D_n					1

(b) Hybrid document similarity

Fig. 4 Representation and similarity of documents using the HYBRID approach

The CONCEPTBASED approach uses the document representation and similarity in Fig. 3. By using the CONCEPTBASED approach we expect to retrieve documents that are similar based on the concepts they contain, and this is obtained from the document-document similarity in Fig. 3b. A standard approach of representing documents would be to define the document similarity based on the term document matrix in Fig. 2b, but this exploits the concept vocabulary only. However, in our approach, we put more emphasis on the domain concepts, so we use the concept document matrix in Fig. 3a, to underpin the similarity between documents.

4.2 The HYBRID *Approach*

The HYBRID approach exploits the relative distribution of the vocabulary in the concept and document spaces to augment the representation of learning resources with a bigger, but focused, vocabulary. So the TF-IDF weight of a term changes depending on its relative frequency in both spaces.

First, the concepts, C_1 to C_m and the documents we wish to represent, D_1 to D_n, are merged to form a corpus. Next, a term-document matrix with TF-IDF weighting is created using all the terms, t_1 to t_T from the vocabulary of the merged corpus as shown in Fig. 4a. For example, entry q_{ik} is the TF-IDF weight of term t_i in D_k. If t_i has a lower relative frequency in the concept space compared to the document space, then the weight q_{ik} is boosted. So, distinctive terms from the concept space will get boosted. Although the overlap of terms from both spaces are useful for altering the term weights, it is valuable to keep all the terms from the document space because this gives us a richer vocabulary. The shaded term vectors for D_1 to D_n in Fig. 4a form a term-document matrix for documents whose term weights have been influenced by the presence of terms from the concept vocabulary.

Finally, the document similarity in Fig. 4b, is generated by computing the cosine similarity between the augmented term vectors for D_1 to D_n. Entry r_{jk} is the cosine similarity of the term vectors for documents, D_j and D_k from Fig. 4a. The Hybrid method exploits the vocabulary in the concept and document spaces to enhance the retrieval of documents.

5 Evaluation

Our methods are evaluated on a collection of topic-labeled learning resources by simulating an e-learning recommendation task. We use a collection from Microsoft Academic Search (MAS) [10], in which the author-defined keywords associated with each paper identifies the topics they contain. The keywords represent what relevance would mean in an e-learning domain and we exploit them for judging document relevance. The papers from MAS act as our e-learning resources, and using a query-by-example scenario, we evaluate the relevance of a retrieved document by considering the overlap of keywords with the query. This evaluation approach allows us to measure the ability of the proposed methods to identify relevant learning resources. The methods compared are:

- ConceptBased represents documents using the domain concepts (Sect. 4.1).
- Hybrid augments the document representation using a contribution of term weights from the concept vocabulary (Sect. 4.2).
- BOW is a standard Information Retrieval method where documents are represented using the terms from the document space only with TF-IDF weighting.

For each of the 3 methods, the documents are first pre-processed by removing English stopwords and applying Porter stemming. Then, after representation, a similarity-based retrieval is employed using cosine similarity.

5.1 Evaluation Method

Evaluations using human evaluators are expensive, so we take advantage of the author-defined keywords for judging the relevance of a document. The keywords are used to define an overlap metric. Given a query document Q with a set of keywords K_Q, and a retrieved document R with its set of keywords K_R, the relevance of R to Q is based on the overlap of K_R with K_Q. The overlap is computed as:

$$Overlap(K_Q, K_R) = \frac{|K_Q \cap K_R|}{min\left(|K_Q|, |K_R|\right)} \tag{1}$$

We decide if a retrieval is relevant by setting an overlap threshold, and if the overlap between K_Q and K_R meets the threshold, then K_R is considered to be relevant.

Our dataset contains 217 Machine Learning and Data Mining papers, each being 2–32 pages in length. A distribution of the keywords per document is shown in Fig. 5,

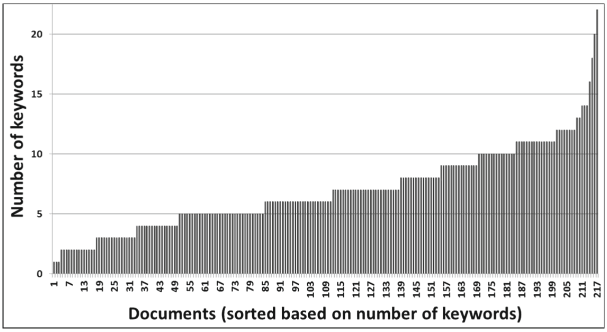


Fig. 5 Number of keywords per Microsoft document

Table 2 Overlap of document-keywords and the proportion of data

Overlap coefficient	Number of pairs	Proportion of data (%)	Overlap threshold
Zero	20251 (86 %)	10	0.14
Non-zero	3185 (14 %)	5	0.25
		1	0.5

where the documents are sorted based on the number of keywords they contain. There are 903 unique keywords, and 1497 keywords in total.

A summary of the overlap scores for all document pairs is shown in Table 2. There are 23436 entries for the 217 document pairs, and 20251 are zero, meaning that there is no overlap in 86 % of the data. So only 14 % of the data have an overlap of keywords, indicating that the distribution of keyword overlap is skewed. There are 10 % of document pairs with overlap scores that are ≥ 0.14, while 5 % are ≥ 0.25.

The higher the overlap threshold, the more demanding is the relevance test. We use 0.14 and 0.25 as thresholds, thus avoiding the extreme values that would allow either very many or few of the documents to be considered as relevant. Our interest is in the topmost documents retrieved, because we want our top recommendations to be relevant. We use precision@n to determine the proportion of relevant documents retrieved:

$$Precision@n = \frac{|retrievedDocuments \cap relevantDocuments|}{n} \tag{2}$$

where, n is the number of documents retrieved each time, *retrievedDocuments* is the set of documents retrieved, and *relevantDocuments* are those documents that are considered to be relevant i.e. have an overlap that is greater than the threshold.

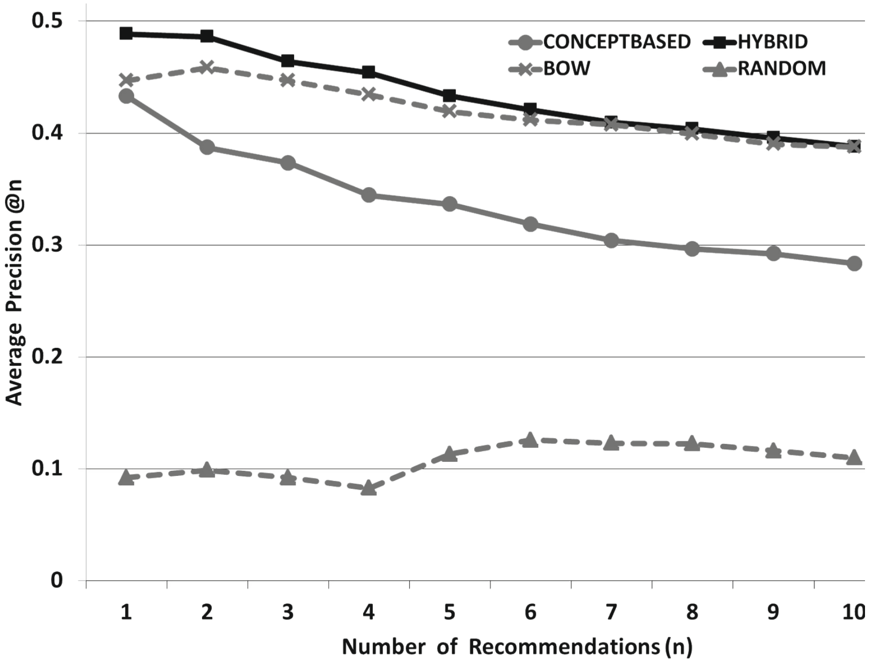


Fig. 6 Precision of the methods at an overlap threshold of 0.14

5.2 Results and Discussion

The methods are evaluated using a leave-one-out retrieval. In Fig. 6, the number of recommendations (n) is shown on the x-axis and the average precision@n is shown on the y-axis. RANDOM (▲) has been included to give an idea of the relationship between the threshold and the precision values. RANDOM results are consistent with the relationship between the threshold and the proportion of data in Table 2.

Overall, HYBRID (■) performs better than BOW (×) and CONCEPTBASED (●), showing that augmenting the representation of documents with a bigger, but focused vocabulary, as done in HYBRID, is a better way of harnessing our background knowledge. BOW also performs well because the document vocabulary is large, but the vocabulary used in CONCEPTBASED may be too limited. All the graphs fall as the number of recommendations, *n* increases. This is expected because the earlier retrievals are more likely to be relevant. However, the overlap of HYBRID and BOW at higher values of n may be because the documents retrieved by both methods are drawn from the same neighbourhoods.

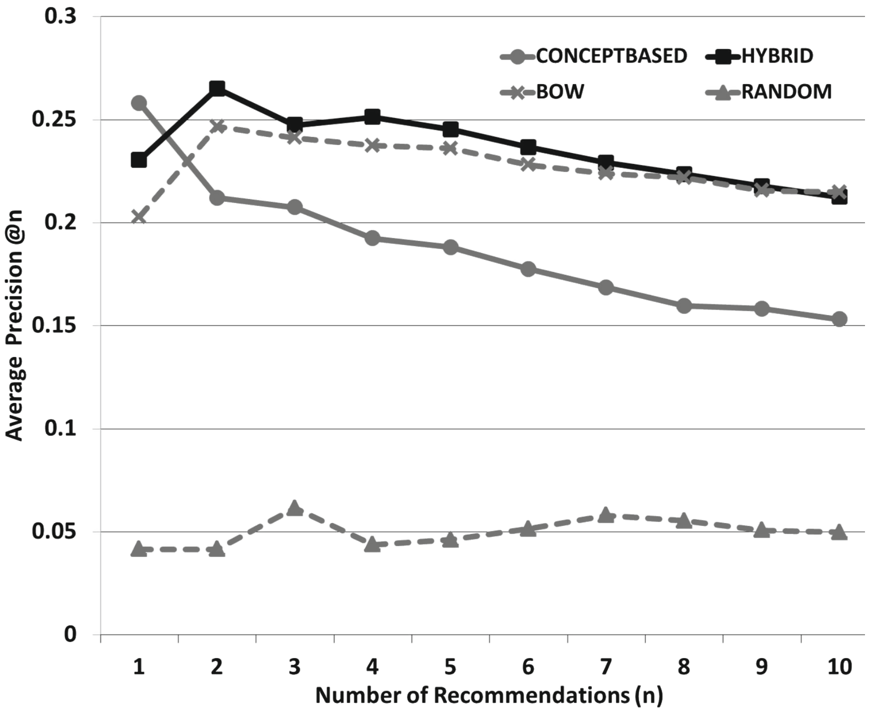


Fig. 7 Precision of the methods at an overlap threshold of 0.25

The relative performance at a threshold of 0.25 in Fig. 7, is similar to the performance at 0.14. However, at this more challenging threshold, HYBRID and BOW do not perform well on the first retrieval. This may be due to the size of the vocabulary used by both methods. Generally, the results show that the HYBRID method is able to identify relevant learning resources by highlighting the domain concepts they contain, and this is important in e-learning. The graphs show that augmenting the representation of learning resources with our background knowledge is beneficial for e-learning recommendation.

6 Conclusions

E-learning recommendation is challenging because the learning resources are often unstructured text, and so are not appropriately indexed for retrieval. One solution is the creation of a concept-aware representation that contains a good coverage of relevant topics. In this paper domain-specific background knowledge is built by exploiting a structured collection of teaching materials as a guide for identifying important concepts. We then enrich the identified concepts with discovered text from an encyclopedia source, and use these pseudo-documents to extend the coverage and richness of our representation.

The background knowledge captures both key topics highlighted by the e-book TOCs that are useful for teaching, and additional vocabulary related to these topics. The concept space provides a vocabulary and focus that is based on teaching materials with provenance. ConceptBased takes advantage of similar distributions of concept terms in the concept and document spaces to define a concept term driven representation. Hybrid exploits differences between distributions of document terms in the concept and document space, in order to boost the influence of terms that are distinctive in a few concepts.

Our results confirm that augmenting the representation of learning resources with our background knowledge in Hybrid improves e-learning recommendation. The larger vocabulary from both concepts and documents has been focused by the use of the vocabulary in the concept space. Although ConceptBased also focuses on the concept space, by using only concept vocabulary, this vocabulary is too restricted for concept-based distinctiveness to be helpful.

In future, the background knowledge will be exploited to support query expansion and refinement in an e-learning environment. One approach would be to represent learners' queries using the vocabulary from our knowledge-rich representation. Alternatively, our background knowledge can be employed to support search by exploration. This would allow learners to search for resources through a guided view of the learning domain.

References

1. Agrawal, R., Chakraborty, S., Gollapudi, S., Kannan, A., Kenthapadi, K.: Quality of textbooks: an empirical study. In: ACM Symposium on Computing for Development, pp. 16:1–16:1 (2012)
2. Beliga, S., Meštrović, A., Martinčić-Ipšić, S.: An overview of graph-based keyword extraction methods and approaches. J. Inf. Organ. Sci. **39**(1), 1–20 (2015)
3. Bousbahi, F., Chorfi, H.: MOOC-Rec: a case based recommender system for MOOCs. Proc. Soc. Behav. Sci. **195**, 1813–1822 (2015)
4. Boyce, S., Pahl, C.: Developing domain ontologies for course content. J. Educ. Technol. Soc. **10**(3), 275–288 (2007)
5. Chen, W., Niu, Z., Zhao, X., Li, Y.: A hybrid recommendation algorithm adapted in e-learning environments. World Wide Web **17**(2), 271–284 (2014)
6. Chen, Z., Liu, B.: Topic modeling using topics from many domains, lifelong learning and big data. In: 31st International Conference on Machine Learning, pp. 703–711 (2014)
7. Coenen, F., Leng, P., Sanderson, R., Wang, Y.J.: Statistical identification of key phrases for text classification. In: Machine Learning and Data Mining in Pattern Recognition, pp. 838–853. Springer (2007)
8. Dietze, S., Yu, H.Q., Giordano, D., Kaldoudi, E., Dovrolis, N., Taibi, D.: Linked education: interlinking educational resources and the web of data. In: 27th Annual ACM Symposium on Applied Computing, pp. 366–371 (2012)
9. Ferrucci, D., Levas, A., Bagchi, S., Gondek, D., Mueller, E.T.: Watson: beyond Jeopardy!. Artif. Intell. **199**, 93–105 (2013)
10. Hands, A.: Microsoft academic search. Tech. Serv. Q. **29**(3), 251–252 (2012)

11. Nasraoui, O., Zhuhadar, L.: Improving recall and precision of a personalized semantic search engine for e-learning. In: 4th International Conference on Digital Society, pp. 216–221. IEEE (2010)
12. Panagiotis, S., Ioannis, P., Christos, G., Achilles, K.: APLe: agents for personalized learning in distance learning. In: 7th International Conference on Computer Supported Education, pp. 37–56. Springer (2016)
13. Porter, M.F.: An algorithm for suffix stripping. Program **14**(3), 130–137 (1980)
14. Qureshi, M.A., O'Riordan, C., Pasi, G.: Exploiting Wikipedia to identify domain-specific key terms/phrases from a short-text collection. In: 5th Italian Information Retrieval Workshop, pp. 63–74 (2014)
15. Ruiz-Iniesta, A., Jimenez-Diaz, G., Gomez-Albarran, M.: A semantically enriched context-aware OER recommendation strategy and its application to a computer science OER repository. IEEE Trans. Educ. **57**(4), 255–260 (2014)
16. Salton, G., Buckley, C.: Term-weighting approaches in automatic text retrieval. Inf. Process. Manag. **24**(5), 513–523 (1988)
17. Völkel, M., Krötzsch, M., Vrandecic, D., Haller, H., Studer, R.: Semantic Wikipedia. In: 15th International Conference on World Wide Web, pp. 585–594. ACM (2006)
18. Witten, I.H., Paynter, G.W., Frank, E., Gutwin, C., Nevill-Manning, C.G.: KEA: Practical automatic keyphrase extraction. In: 4th ACM Conference on Digital libraries, pp. 254–255 (1999)
19. Yang, H.L., Lai, C.Y.: Motivations of Wikipedia content contributors. Comput. Hum. Behav. **26**(6), 1377–1383 (2010)
20. Yang, K., Chen, Z., Cai, Y., Huang, D., Leung, H.: Improved automatic keyword extraction given more semantic knowledge. In: International Conference on Database Systems for Advanced Applications, pp. 112–125. Springer (2016)
21. Zhang, X., Liu, J., Cole, M.: Task topic knowledge vs. background domain knowledge: impact of two types of knowledge on user search performance. In: Advances in Information Systems and Technologies, pp. 179–191. Springer (2013)

Knowledge Discovery and Data Mining