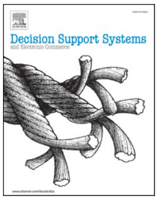
Contents lists available at ScienceDirect

Decision Support Systems

journal homepage: www.elsevier.com/locate/dss



Methodology for refining subject terms and supporting subject indexing with taxonomy: A case study of the APO digital repository

Yong-Bin Kang [a,*], Jihoon Woo [a], Les Kneebone [b], Timos Sellis [c,1]

[a] Swinburne University of Technology, Australia
[b] Analysis & Policy Observatory, Australia
[c] Facebook, USA

ARTICLE INFO

Keywords:
Controlled terms
Subject taxonomies
Subject headings
Thesaurus construction
APO
Digital library

ABSTRACT

In digital repositories, it is crucial to refine existing subject terms and exploit a taxonomy with subject terms, in order to promote information retrieval tasks such as indexing, cataloging and searching of digital documents. In this paper, we address how to refine an existing set of subject terms, often containing irrelevant ones or creating noise, that are used to index digital documents. Further, we present how to automatically induce a subject term taxonomy to capture and utilise the semantic relations among subject terms. Most related works have little studied these problems, focusing mostly on creating subject terms or building a taxonomy of key terms from text documents. We propose a methodology[2] for refining an existing set of subject terms in a digital repository by identifying their semantics, as well as inducing a taxonomy with subject terms by analysing their mutual usages, maximising their semantic relatedness. Then, we present a case study using the (Analysis & Policy Observatory) APO digital repository to analyse the proposed methodology and demonstrate its applicability. Further, to validate the generalisability of the proposed taxonomy inducing method, we evaluate it using a gold-standard taxonomy in life sciences, Medical Subject Headings (MeSH), in comparison with the state–of-the-art taxonomy inducing method, TaxoFinder. Our evaluation shows that our methodology has a high potential for refining an existing set of subject terms and capturing their semantic relationships by inducing a subject term taxonomy.

1. Introduction

Digital repositories (or libraries[3]) have emerged as essential information systems that serve repositories of digital documents as well as provide search and retrieval mechanisms via user interaction [1]. Advances in information retrieval research have significantly enhanced the access, functionality and technical capabilities of digital repositories. In a digital repository, *subject terms* are an essential block for descriptive cataloging and indexing documents [2]. These terms are usually derived from some type of controlled vocabulary such as predefined keywords associated with the underlying documents [3]. Thus, subject terms are considered as a key asset in a digital repository, and these terms significantly contribute not only to describing information or knowledge pieces of digital documents but also to improving the relevance of search results [4]. As a result, to make effective use of digital documents, developing and utilising useful subject terms plays a crucial role in determining the quality of a digital repository.

Creating a taxonomy of subject terms is another key to promoting information retrieval tasks such as indexing, cataloging and searching of information from digital documents [5–7]. In digital repositories, we consider three benefits of using a subject term taxonomy. First, the ability to index digital documents can be improved, regardless of indexing methods (i.e., manually, semi-automatically or automatically), by utilising semantic associations between the subject terms induced from the taxonomy. Second, using semantic knowledge about the taxonomy provides better understanding about the underlying subject terms for humans, thereby facilitating their refinement (i.e., the improvement or clarification of subject terms). Third, a subject term

[*] Corresponding author at: Department of Media and Communication, Swinburne University of Technology, John St, Hawthorn, VIC 3122, Australia.
E-mail addresses: ykang@swin.edu.au (Y.-B. Kang), jihoonwoo@swin.edu.au (J. Woo), lkneebone@apo.org.au (L. Kneebone), tsellis@fb.com (T. Sellis).

[1] Work done while at Swinburne University of Technology.

[2] The source code of the proposed methodology is available at https://github.com/Yongbinkang/SubjectTracker

[3] The boundaries between digital repositories and digital libraries are sometimes blurred, as repositories also host digital collections for access [1]. In this paper, we do not distinguish these terms and interchangeably use them.

https://doi.org/10.1016/j.dss.2021.113542
Received 13 October 2020; Received in revised form 17 February 2021; Accepted 5 March 2021
Available online 13 March 2021

taxonomy can improve researchers or end-users to search documents by linking and suggesting related subject terms and by offering their hierarchical structure that helps to navigate about them more easily.

In this paper, we tackle two challenging problems. The first is to refine an *existing* set of subject terms that often contain irrelevant ones in a digital repository. We refer to an irrelevant subject term as the subject term that may not be useful for indexing the document collection in the target repository. Overall, this term is rarely used to index the collection and can cause confusion for indexing in a digital repository. The second is to automatically *induce a taxonomy* from refined subject terms to capture their semantic relations. Our research motivation is two-fold. First, expediting the reuse of existing subject terms has been highlighted as a key to maximising their value [8]. In some digital repositories, a controlled vocabulary of subject terms is often *readily available*. For instance, Analysis & Policy Observatory (APO),[4] the largest open access repository for public policy and research literature in Australia, has already used a combination of some portion of subject terms drawn from a general-use controlled vocabulary, Faceted Application of Subject Terminology (FAST),[5] and the subject terms that the repository curators[6] have manually identified. However, as highlighted in [3], there often exists many irrelevant subject terms in the underlying repository. Further, in order to identify relevant subject terms from a general-use controlled vocabulary, the success of this task requires a high level of comprehensiveness of the underlying documents and the semantic coverage of the terms in such a vocabulary [9]. However, this task may not be obvious and thus can be very difficult to achieve. As the second facet of our motivation, inducing a taxonomy from subject terms has been relatively overlooked. This is a challenging problem, as these terms themselves do not contain explicit relationships from which a taxonomy can be constructed.

Most existing approaches have little focused on the challenging problems mentioned above. First, these approaches have mainly attempted to propose methods for creating subject terms (or keywords). Instead, our focus is to automatically refine an existing set of the subject terms by analysing their semantics. Second, the related works have mostly focused on automatically building a taxonomy of important terms from text documents [5,10]. However, we focus on inducing a subject term taxonomy from refined subject terms by analysing their mutual usages.

This paper makes three main contributions: First, we propose a method for refining an existing set of subject terms $\mathscr{S}$ that possibly contain irrelevant ones in a given document collection $\mathscr{D}$ (Section 3). Given $\mathscr{S}$, our refinement process takes two steps: (1) for each document $d \in \mathscr{D}$, we identify additional subject term candidates that may be relevant but not previously assigned to d. Such candidates are added to $\mathscr{S}$ (Section 3.1); and (2) we filter out irrelevant subject terms from $\mathscr{S}$ and merge insignificant subject terms with more significant ones based on their similarities to produce a more precise set of subject terms, i.e., called refined subject terms $\mathscr{S}'$, to improve their semantic coverage (Section 3.2). Second, we propose an approach for inducing a taxonomy from $\mathscr{S}'$ by integrating their mutual usages for indexing documents and their semantics (Section 4). For this, we apply the *subsumption method* [11], with our proposed objective function that maximises the semantic relatedness of subject terms in the induce taxonomy. Third, we propose a case study using the APO repository to show the applicability of the proposed methodology (Section 5). Further, to show the generalisability of the proposed taxonomy inducing approach, we evaluate its effectiveness using MeSH,[7] in comparison with the state–of-the-art taxonomy inducing method, TaxoFinder [5] (Section 6). Our case study and evaluation show that the proposed methodology has high potential to be used for refining an existing set of subject terms and capturing their semantic relationships by inducing their taxonomy.

This paper is organised as follows. Section 2 provides related works and background of the APO repository. Section 3 presents an overview of the proposed methodology and describes the process for refining subject terms. Section 4 discusses our approach for inducing a subject term taxonomy. Section 5 shows a case study of our proposed methodology using the APO repository. Section 6 presents our evaluation of our approach for inducing a subject term taxonomy to show its generalisability. Section 7 presents the conclusion of this paper.

2. Related work and background

In this section, first, we present how information is organised using subject terms and taxonomies in digital repositories. Then, we review research works that focused on creating or refining subject terms (or keywords) in the community of information and library science. Afterwards, we discuss related works on inducing a taxonomy from a knowledgebase or subject terms. Finally, we introduce the APO digital repository for our case study.

2.1. Information organisation through subject terms

In digital repositories, a controlled vocabulary of subject terms is a list of terms or phrases used for descriptive cataloging, tagging or indexing [12]. The term ‘controlled’ means that such terms can be typically used under specific conditions, and also changed by a controlled vocabulary editor, metadata creator or taxonomist [12]. In digital repositories, metadata indicates information about digital data that help to make quick access to it. Metadata can enhance the process of resource (e.g., documents) discovery by disclosing specific information about the given resource. Some general forms of metadata about documents can be seen as title, author and publisher. Subject terms (or headings) are integral part of metadata and specific examples of content-based metadata [13]. These terms mainly describe the content of each document in a digital repository.

A taxonomy is a hierarchical classification of a set of things or concepts in a domain. Building (or inducing) a taxonomy is an essential task for knowledge acquisition, sharing and classification in various domains [5]. In digital libraries, taxonomies have been widely used to organise collections in a digital repository [14]). For example, a multiple-disciplinary repository arXiv uses a certain taxonomy relevant to each discipline to classify the manuscripts submitted under the discipline (e.g., computer science uses ACM classification[8]). MEDLINE®,[9] one of the largest databases for life sciences, is indexed by 29k+ MeSH terms organised by a hierarchical structure. A public digital library, Digital Public Library of America (DPLA), provides access to Americans' most trusted sources of shared digital materials from libraries, archives, and museums around the world.[10] Europeana is a web portal covering over 10 million cultural and scientific artefacts from European museums. Europeana uses the Art & Architecture Thesaurus (AT&T) to describe

[4] https://apo.org.au

[5] FAST is a general-use controlled vocabulary (https://www.oclc.org/research/themes/data-science/fast.html) based on the Library of Congress Subject Headings (https://id.loc.gov/authorities/subjects.html).

[6] In this paper, a curator is referred to as a person who manages, administers, or organises information in a digital repository, tagging a document with relevant subject terms. An indexer (also called tagger) is a person who does the part of the role of a curator, specifically focusing on assigning subject terms. To simplify our presentation, we use the term curator to indicate the one who assigns subject terms.

[7] http://www.nlm.nih.gov/mesh

[8] https://www.acm.org/publications/computing-classification-system/how-to-use

[9] https://www.nlm.nih.gov/bsd/medline.html

[10] https://en.wikipedia.org/wiki/Digital_Public_Library_of_America

cultural artefacts.[11] A survey of other readily available taxonomies have been used in digital repositories [15].

Subject indexing is the task of describing the subjects of a document. This task focuses on assigning predefined subject terms to a document to indicate what the document is about. Also, this task yields important information as retrieval of information depends to a large extent on the quality of indexing [4]. Typically, subject indexing has been made manually, i.e., assigned by a curator according to their content or aboutness [16]. However, manual subject indexing causes a challenge that requires huge time and efforts of curators. Also, if there are multiple curators in a digital repository and each curator has a different level of expertise, the indexing task is likely to produce inter-rater disagreements and inconsistency. To compensate this problem and due to the increasing volume of information in digital formats, there have been many studies for automatic subject indexing. For example, the work [2] formulated subject indexing as a text classification problem and used the Support Vector Machine classifier to automatically classify documents with subject terms. Helping Interdisciplinary Vocabulary Engineering (HIVE) [17] is a representative system that provides a machine learning approach for automatic subject indexing. HIVE supports an integration of multiple general, domain-specific vocabularies to aid with metadata generation. HIVE uses the KEA++ library [13] for subject indexing. Automatically indexing documents with subject terms is still challenging and is a hot research topic in machine learning.

2.2. Creating and refining subject terms

Creating high quality subject terms is essential for organising and making accessible the growing number of digital documents in a digital repository. As highlighted in [18], manual-based approaches were often used to create subject terms. However, these approaches usually require labor intensive time and cost. Also, these approaches generally raise other issues such as producing noise, irrelevant, subjective and inconsistent subject terms [19]. It was also emphasised that a semi-automatic subject term generation method can be useful using extra information resources (e.g., Web), to complement the drawbacks of the manual-based work [20]. Due to the advances of natural language processing (NLP) and machine learning techniques, automatically identifying subject terms or keywords were extensively studied in information science. For example, the study [13] proposed a method that selects candidate subject terms and then ranks them by their significance based on their properties such as statistical, semantic, and encyclopedic knowledge. These properties were combined using a machine learning algorithm that models human indexing behavior from examples. However, the ability to identify relevant subject terms still remains as a challenge as their recall values are relatively low. For example, CFinder [21] proposed a concept extraction method and applied it for extracting subject terms in a mass gathering corpus, but its accuracy only reached to around 50% in terms of F1-measure in which the evaluation was done based on human experts. Also, a recent keyword extraction method, YAKE! [22] showed that its accuracy, in terms of F1-measure using comprehensive text datasets, is less than 50%.

In this paper, our focus differs from the above works in that we tackle the problem of refining subject terms from an existing set of subject terms that may contain noise and irrelevant terms.

2.3. Inducing a taxonomy of subject terms

Recent growth of NLP techniques has boosted the development of automated methods for building a taxonomy from a document collection. Some earlier works mainly focused on analysing linguistic patterns of concepts (i.e., important terms) that appear in a given document collection. For example, the works [23,24] are representative works that used predefined lexico-syntactic patterns (e.g., A is the same as/know as/call/refer to as B) to extract concepts and their taxonomic relations, to induce a concept taxonomy. In general, the quality of the lexico-syntactic based approaches mainly depends on their ability to define the grammatical functions of terms in sentences. Some limitations of these works include that lexico-syntactic patterns may not frequently appear in the underlying documents. Thus, usually, predefining such patterns is very difficult and requires additional investigation on extra knowledge sources to discover taxonomic relations of concepts. TaxoFinder [5] used a graph-based approach for building a taxonomy by identifying concepts (i.e., domain-specific keywords) from a document collection. It used a method for building a concept graph representing how such concepts are associated together based on their co-occurrences. Then, it applied a graph analytic method (i.e., finding the Maximum Spanning Tree) to induce a taxonomy from the concept graph, exploiting associative strengths among the concepts. The *subsumption method* focused on the co-occurrences of concepts to induce taxonomic relations of concepts [11]. It relies on the idea that a concept A subsumes another concept B (i.e., A is the hypernym of B), if the documents where B appears are a subset of the documents that A appears. The common intuition of the above works for inducing a taxonomy is to use the co-occurrence based statistical analysis of terms (or concepts). If two terms more frequently appear together, there is a higher probability that these terms are semantically related. In our work, we adopt the subsumption method [11] for inducing a taxonomy from predefined subject terms, due to its proven performance and a faster creation of taxonomic relations.

Automatically inducing a taxonomy from subject terms associated with a given document collection has received little attention, in comparison with the studies of inducing a taxonomy of important terms from a document collection. This is a challenging task as predefined subject terms themselves do not contain explicit relationships from which a taxonomy can be constructed. We have found one similar study [25] to our work in that it proposed methods for automatically inducing a taxonomy from predefined keywords based on the exploitation of external 'knowledge' and 'context' of the keywords. As the source of such knowledge and context, [25] used Probase [26], which is constructed by automatic ontology population, and a search engine (i.e., throwing each keyword into a search engine, and aggregating the keywords from its top k search result), respectively, From another angle, clustering approaches have been used to automatically build a taxonomy from predefined keywords. In these approaches, the implied idea is to assign such keywords into clusters in the way that keywords in the same cluster are more similar to each other in some sense than those in other clusters [27]. The premise in the clustering approaches is that there would be a good similarity measure that can capture the semantics between keywords.

Our approach for inducing a subject term taxonomy differs from the above works in that our idea for inducing a taxonomy from subject terms is to leverage their past mutual usages (i.e., co-occurrences) for indexing the given document collection. Leveraging such usage information has also been proven to be effective in inducing a taxonomy [11].

2.4. Analysis & policy observatory (APO) repository

APO is a multi-domain digital repository that provides a grey literature collection "comprised of research and information resources produced and disseminated... by organisations, outside of the commercial or scholarly publishing industry focusing on public policy and research" [28]. Grey literature refers to an extensive and complex source of text-based information, usually produced by government, academics, business and industry but not controlled by commercial publishers [29]. APO curates and indexes its documents with subject terms. APO uses a combination of a third-party set of the subject terms which are a portion of the FAST terms and locally built subject terms by the APO curators.

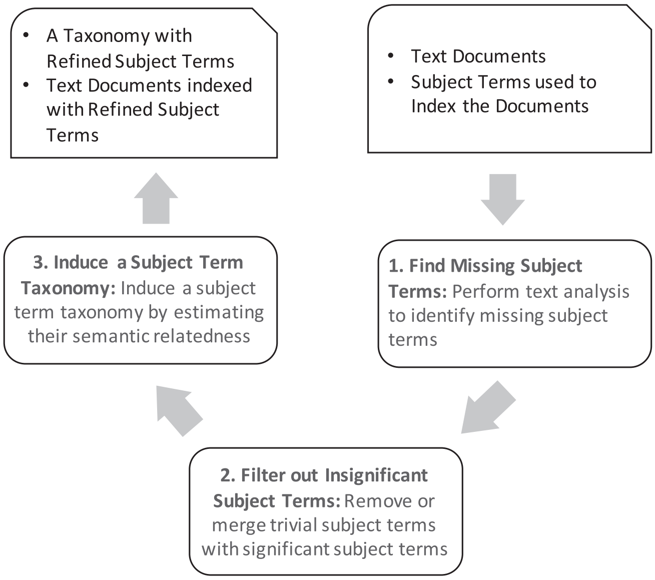
[11] https://pro.europeana.eu/post/europeana-enriches-its-data-with-the-art-and-architecture-thesau

FAST is a vocabulary of subject terms derived from the Library of Congress Subject Headings (LCSH) that is one of the largest subject vocabularies used by digital libraries. FAST is comprised of a 9-facet vocabulary (i.e Personal names, Corporate names, Meeting names, Geographic names, Events, Titles, Time periods, Topics, and Form/Genre) with approximately 1.8 million subject terms across all facets. A number of agencies and institutes adopt FAST for a variety of purposes, e.g., British Library and National Library of New Zealand, for indexing digital materials [30,31]. The FAST terms include a general, broad range of subject terms thereby these may be fitting well into a general-purpose repository.

Currently, APO has collected a controlled vocabulary of around 5700 subject terms over a number of years to index its document collection. There are some issues in facilitating these terms to make effective searching in APO. First, there is a high number of irrelevant and non-reusable subject terms in the controlled vocabulary [3]. One reason is that many subject terms were added as part of a bulk metadata import project. Another reason is that the APO curators sometimes had difficulties to review and choose relevant subject terms from the existing ones. Due to the above reasons, it was also easy to inadvertently create new subject terms that overlapped semantically with the existing subject terms. Second, the APO subject terms could not be navigated due to a lack of subject term reference structure. That is, as a subject term taxonomy had not been established, the APO curators had difficulties to discover broader or narrower semantics of subject terms, especially when adding new subject terms. In this paper, we tackle and address the above challenges by proposing a methodology for refining an existing set of subject terms as the case in APO, and inducing a subject term taxonomy.

3. Methodology for refining subject terms

In this section, we present the details of our proposed methodology for refining subject terms, followed by the method for inducing a subject term taxonomy in Section 4. The overview of the methodology comprised of the three steps is depicted in Fig. 1. First, given the text documents $\mathscr{D}$ in a repository and existing set of subject terms $\mathscr{S}$ used to index $\mathscr{D}$, we identify the subject terms that are potentially relevant but previously not assigned to related documents in $\mathscr{D}$ (Section 3.1). Our premise behind in this step is that indexing some documents in $\mathscr{D}$ with $\mathscr{S}$ are likely to be inconsistent and inaccurate due to human errors, a lack of humans' understanding of the semantics of some documents in $\mathscr{D}$ and the subject terms, or a machine's inefficiency as discussed in Sections 2.2. Second, to make a more condensed and meaningful set of the subject terms, we filter out irrelevant subject terms by removing or merging them with more significant subject terms (Section 3.2). Third, using the refined subject terms $\mathscr{S}'$ after performing the second step, we induce a taxonomy that classifies $\mathscr{S}'$ according their semantic relationships (Section 4). In the following, we present more detailed descriptions about each step.



• A Taxonomy with Refined Subject Terms
• Text Documents indexed with Refined Subject Terms
• Text Documents
• Subject Terms used to Index the Documents
3. Induce a Subject Term Taxonomy: Induce a subject term taxonomy by estimating their semantic relatedness
1. Find Missing Subject Terms: Perform text analysis to identify missing subject terms
2. Filter out Insignificant Subject Terms: Remove or merge trivial subject terms with significant subject terms

Fig. 1. The overview of the proposed methodology comprised of the three steps to induce a subject term taxonomy.

3.1. *Identification of missing subject terms*

One key to finding accurate information from the document collection $\mathscr{D}$ in a digital repository is whether each document has been indexed with a relevant set of subject terms. With this in mind, our goal here is to identify *missing subject terms* for $\mathscr{D}$. Given a document $d \in \mathscr{D}$, a missing subject term is referred to as a potentially useful subject term candidate for indexing d but previously not assigned as a subject term to d. This term may reflect some portion of the actual content of d.

We perform the following procedure to identify missing subject terms by a string matching technique. Our assumption is that if a subject term $s \in \mathscr{S}$ appears in a document $d \in \mathscr{D}$, s can be a subject term candidate for d. Thus, we check if d contains each term $s \in \mathscr{S}$ in its text content. However, a simple string matching that scans d and finds the strings that exactly match s cannot work properly. To illustrate, consider the following example paragraph:

- "RMIT University undertook the research with a XXX Innovation Research Grant. This report presents insight into the complexity and multiplicity of place based experiences of **social exclusion**. **IT** has been significantly developed over the last decade. It is reported that indigenous engagement with vocational education and training (**VET**) has improved significantly."

Suppose that the following subject terms exist in $\mathscr{S}$: 'social exclusion', 'VET', and 'IT' which are denoted in boldface in the text. Then, we need to consider these issues to identify them. First, if we were to apply an *exact* matching technique that attempts to find the strings that exactly match a subject term s, a problem arises. For example, given 'social exclusion' and 'VET', we cannot identify these terms as the former is concatenated with punctuation '.' (i.e., 'social exclusion.'), and the latter is concatenated with '(' and ')' (i.e., '(VET)'). Second, if we attempt to solve the above problem by an *inclusion* matching technique that identifies the strings that contain a given subject term, we can solve the above problem, however, another problem arises. For example, the term 'RMIT' can be incorrectly retrieved given a subject term 'IT', as 'IT' is included in 'RMIT'. Third, regardless of either the exact or inclusion matching technique, if we do not distinguish uppercase and lowercase characters, we could retrieve incorrect terms. For example, we could retrieve the term 'It', given a subject term 'IT', as they have the same characters.

In our approach, we address the above three problems in the following ways. To address the first problem, we remove all special characters (i.e., non-alphanumeric characters, or numeric characters) in the content of all documents $\mathscr{D}$. To address the second problem, we distinguish whitespaces before and after a word in each sentence in our matching. By doing so, we cannot retrieve the term 'RMIT' given a subject term 'IT', as we find a string ' IT ' not 'IT'. To address the third problem, we distinguish uppercase and lowercase characters. So we do not retrieve 'It' accidentally as 'It' differs from 'IT' in capitalisation. Finally, the identified missing subject terms are added to $\mathscr{S}$.

3.2. *Filtering out irrelevant subject terms*

Although we have identified missing subject terms, all subject terms

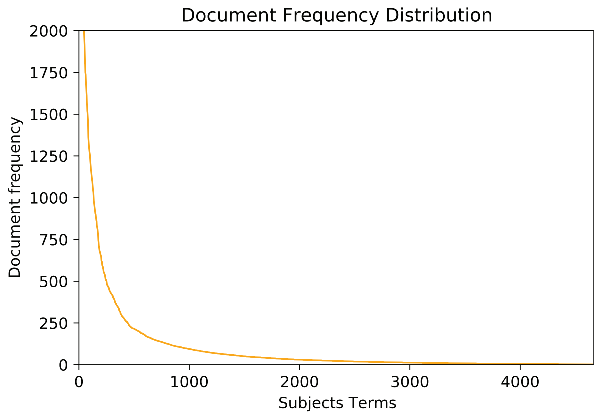


Fig. 2. Frequency distribution of the subject terms in the APO corpus.

in $\mathscr{S}$ may not be relevant for indexing the document collection $\mathscr{D}$. To filter out irrelevant subject terms from $\mathscr{S}$, our approach is to use their indexing usage in terms of their frequencies. There can be a high imbalance in the usage of subject terms $\mathscr{S}$ over $\mathscr{D}$, where some subject terms have been dominantly used but some others have rarely used for indexing $\mathscr{D}$. This situation can cause the following problems. First, the subject terms that are rarely used can unnecessarily increase the vocabulary size of subject terms and this can make it hard for curators (also machines) to choose relevant subject terms for indexing a new document. Second, given similar documents, different curators may assign different subject terms to these documents thus increasing inconsistency and inaccuracy. Third, if there are similar subject terms, sometimes, curators may not clearly determine which ones need to be used to index a given document. For example, Fig. 2 shows the document frequency distribution of the subject terms used for indexing the entire document collection in the APO repository. The x-axis shows the number of the subject terms sorted by their document frequencies, and the y-axis shows their document frequencies. The document frequency of a subject term s indicates the number of times s is used to index the documents in a given repository. As we can see, the imbalance ratio of the document frequencies of the subject terms is very high, where some terms are dominantly (very frequently) used (e.g., see the subject terms < 500 on the x-axis) but the majority of the subject terms are rarely (very infrequently) used (e.g., see the subject terms > 1000 on the x-axis). Also, as observed, there is a very long tail of the subject terms that are rarely used.

To address the above three problems, our solution is to reduce the vocabulary size of subject terms, while minimising the loss of the semantics of subject terms in $\mathscr{S}$. The detailed procedure is presented as follows. First, we define *primary subject terms* $\mathscr{S}_p \subseteq \mathscr{S}$ that are dominant subject terms for indexing $\mathscr{D}$, where their total document frequencies are greater than or equal to the minimum document frequency, `min-df`. The rest of the subject terms are classified as *secondary subject terms* $\mathscr{S}_s \subset \mathscr{S}$. Thus, $\mathscr{S}_p \cup \mathscr{S}_s = \mathscr{S}$. The idea of selecting `min-df` is restrictive, capable of filtering rarely-used subject terms. For this, we find the distribution of document frequencies of $\mathscr{S}$ (i.e., Fig. 2). Second, we determine whether each $s \in \mathscr{S}_s$ can be relevant or not. For this, our approach is to measure the similarity between s and each term $s^{'} \in \mathscr{S}_p$. If we find s' highly similar to s (i.e., if the similarity is greater than or equal to the minimum similarity threshold, denoted as `min-sim-syns`), we add s into the *synonym* list of s'. However, if there is no such term s', we simply remove s from $\mathscr{S}$ based on our premise that this would be an irrelevant subject term. In this paper, given a subject term, its similar subject term or synonym is defined as a subject term that has the same or nearly same semantics (or meaning) but differs in lexical representation. Thus, synonyms share a common set of contexts in which they are mutually exchangeable (e.g., bill and invoice, goods and products).

Below we elaborate our two approaches for measuring the similarity between s and s' according to their word length. Note that all the similarity scores in these approaches are normalised to real numbers between 0 (completely dissimilar) and 1 (identical).

3.2.1. Similarity estimation for single-word subject terms

We note that some subject terms in $\mathscr{S}_s$ are single-words whose word length is 1 (e.g., man). Our first approach for finding a similar primary subject term given a single-word subject term $s \in \mathscr{S}_s$ takes two steps. In the first step, we measure the similarity between s and each single-word subject term $s^{'} \in \mathscr{S}_p$ based on their *general semantics* obtained from WordNet.[12] WordNet mimics human logics focusing on word senses and connections, and is a large lexical database of English words. In WordNet, the set of synonyms are grouped together into sets of cognitive synonyms, and these sets are linked together based on their semantic similarities and lexical relations. Given s (e.g., man), if there is a term s' (e.g., male) and their similarity score is $\geq$ `min-sim-syns`, s is added into the synonym list of s'. To measure the similarity, we use the well-known similarity measure proposed by Jiang and Conrath [32].

In the second step, given s, if we cannot find any single-word subject term s' where their similarity is $\geq$ `min-sim-syns`, we estimate their similarity using *word embeddings*. Our aim here is to additionally uncover their semantic relations, whose senses are not found in WordNet, using word embeddings based on their co-occurrences in $\mathscr{D}$. Using word embeddings, we can find synonyms for s by using its nearest neighbors that appear in similar contexts with s. Word embedding is a feature learning technique in NLP that can effectively capture semantic and syntactic word similarities from a document collection. Word2vec [33] is one of the models that generate such a mapping using an artificial neural network that is trained to reconstruct linguistic contexts of words. As the input, Word2vec takes a document collection (i.e., $\mathscr{D}$) and as the output, it produces a vector space where similar words are positioned close to one another. It has been well demonstrated that Word2vec has many advantages for analysing semantic analysis of words [34]. More specifically, we build a word embedding model $E_{\mathscr{D}}$ from $\mathscr{D}$. Then, given a single-word subject term $s \in \mathscr{S}_s$ and each $s^{'} \in \mathscr{S}_p$, we measure their similarity $sim(s, s')$ as follows:

$$sim(s, s^{'}) = cos_{E_{\mathscr{D}}}(\mathbf{s}, \mathbf{s}^{'}), \tag{1}$$

where $cos_{E_{\mathscr{D}}}(\mathbf{s}, \mathbf{s}^{'})$ is the cosine similarity using $E_{\mathscr{D}}$. The cosine similarity is given in Eq. (2). $\mathbf{s}$ and $\mathbf{s}'$ are referred to as the embedded vectors of s and s', respectively. Based on this equation, we add s to the synonym list of s' if the similarity is $\geq$ `min-sim-syns`, otherwise, we remove s from $\mathscr{S}$.

In both steps, if there are more than two primary subject terms that are similar to s, we compare their similarity scores and choose the one with the highest score. Also note that in the second step, the most similar s' can be either a single-word or a multi-word subject term. Further, we propose to use the same threshold `min-sim-syns` in both steps, to identify s' using the same threshold degree of similarity. All the similarity measures in this paper tend to be 1 as two terms being compared have more and more common 'characteristics', although the notion of each measure is subjective. Thus, by using the same threshold degree of similarity, `min-sim-syns`, we aim to facilitate comparative analysis of subject terms in a consistent manner.

For the second step, we build a Word2vec model $E_{\mathscr{D}}$ from $\mathscr{D}$, after removing stopwords and applying lemmatization, in order to learn embeddings of terms that appear in $\mathscr{D}$. As the output, each term in $E_{\mathscr{D}}$ is represented by a *numerical vector* so that we measure the similarity between terms using the cosine similarity between their corresponding vectors. The cosine similarity in Eq. (1) is measured by:

[12] https://wordnet.princeton.edu

$$cos_{E_{\mathscr{D}}}(\mathbf{s}\mathbf{s}^{'}) = \frac{\mathbf{s}\cdot\mathbf{s}^{'}}{\parallel \mathbf{s} \parallel \parallel \mathbf{s}^{'} \parallel} = \frac{\sum_{i=1}^{n} s_i s_i^{'}}{\sqrt{\sum_{i=1}^{n}(s_i)^2}\sqrt{\sum_{i=1}^{n}(s_i^{'})^2}}, \tag{2}$$

where $\mathbf{s}$ and $\mathbf{s}'$ are n-dimensional vectors, and s_i and s_i' denote the i-th numerical value of vectors $\mathbf{s}$ and $\mathbf{s}'$, respectively. Since the cosine similarity produces scores between -1 and 1, we convert the negative scores to zero. Lastly, one issue to be clarified is how to define the vector $\mathbf{s}'$, as s' can be comprised of multi-terms (e.g., 'community health'). Formally, suppose that s' consists of an ordered set of n-terms, $s' = (w_1, \ldots, w_n)$. Then, the vector $\mathbf{s}'$ is estimated by the average of the embedded vectors of its constituent terms,

$$\mathbf{s}^{'} \approx \frac{1}{n}(\mathbf{w_1} + \ldots + \mathbf{w_n}), \tag{3}$$

where $(\mathbf{w}_1, \ldots, \mathbf{w}_n)$ are the embedded vectors of $(w_1, \ldots, w_n)$, respectively.

In this approach, to identify the most similar subject term $s^{'} \in \mathscr{S}_p$ given a single-word subject term $s \in \mathscr{S}_s$, we emphasise that we exploit both human-created logics about synonyms derived from WordNet and a machine learning model Word2vec together to bring their merits into the calculation of similarities between s and s'. By doing so, our aim is to more thoroughly measure the similarity between s and s' than a single measure only.

3.2.2. *Similarity estimation for multi-word subject terms*

As the second approach for measuring the similarity between s and s', some subject terms in $\mathscr{S}_s$ can be multi-words whose word length is greater than 1 (e.g., 'social science'). Here, a question is: how can we measure the similarity between a multi-word secondary subject term and a primary subject term? Using WordNet cannot be a good idea for this purpose, as WordNet is only useful for measuring the similarity between single-words whose senses are found in WordNet. Our approach is to consider the semantics of multi-word subject terms for estimating the similarity using word embeddings. Using the same word embedding model $E_{\mathscr{D}}$ built in Section 3.2.1, given a multi-word $s \in \mathscr{S}_s$ and a primary subject term $s^{'} \in \mathscr{S}_p$, we measure their similarity $sim(s, s')$ using Eq. (1). The cosine similarity in Eq. (1) between the embedded vectors $\mathbf{s}$ and $\mathbf{s}'$ is measured using Eq. (2). To calculate Eq. (2) in Section 3.2.1, we have applied Eq. (3) to the multi-word primary subject terms in $\mathscr{S}_p$. However, we now need to apply Eq. (3) to both $\mathbf{s}$ and $\mathbf{s}'$, as both can be multi-words.

In this approach, as done in Section 3.2.1, we add s to the synonym list of s' if the similarity is $\geq$ `min-sim-syns`, otherwise, we remove s from $\mathscr{S}$. Also, if there are multiple primary subject terms similar to s, we choose the one with the highest score.

4. Inducing subject term taxonomy

In Section 3, we have discussed the process for refining existing subject terms. The outcome of the primary subject terms are called *refined subject terms* $\mathscr{S}^{'}$, where each one is associated with its synonyms. In the rest of this section, to simplify our presentation, subject terms are referred to $\mathscr{S}^{'}$.

Although we have generated $\mathscr{S}^{'}$, we may still have a difficulty for understanding and representing the semantic relatedness of terms in $\mathscr{S}^{'}$. Specifically, there are three important questions to be addressed. First, how can we conceptualise the semantic relatedness of terms in $\mathscr{S}^{'}$ in a formal way to enhance their classification? Second, whether there is a better way to easily interpret their semantic relationships? Third, how can we leverage the semantic relationships among terms in $\mathscr{S}^{'}$ to enhance indexing and searching capability for the document collection $\mathscr{D}$? To address them, our solution is to automatically induce a taxonomy of subject terms $\mathscr{S}^{'}$. Such a taxonomy can enable human indexers or machines to identify the semantic structure of terms in $\mathscr{S}^{'}$ by navigating their relationships in the taxonomy. Fig. 3 shows an example taxonomy consisting of nine social-related subject terms. Each arrow shows a taxonomic relation (subsumer-subsumee or ancestor-descendent). The key relationship in a taxonomy is, namely, 'is-a' relationship. Its important nature lies in that its structure is hierarchical and thus its transitivity is logically inferred by navigating the hierarchical relationships between concepts (i.e., subject terms). The lower concepts inherit all the characteristics from their ancestor concepts. Namely, the higher the position of a concept, the more abstract it is. In addition, highly related concepts are grouped together and the path between two different concepts in the taxonomy reflects how these are semantically related. In Fig. 3, we see that 'Social' is the most abstract term whose immediate descendants are 'Social Policy', 'Social Inclusion', and 'Social Work'. Also, we see that there are five most specific terms that have the common ancestor 'Social'. In addition, intuitively, we can say that the similarity between 'Social' and 'Social Policy' is higher than the similarity between 'Social' and 'Australian Social Policy'.

To induce a taxonomy from $\mathscr{S}^{'}$, our approach is to use the subsumption method [11]. Its fundamental is to use the co-occurrences of subject terms for indexing each document in $\mathscr{D}$. From the co-occurrence knowledge, we can induce that a subject term A *subsumes* another subject term B (i.e., A is the hypernym of B) if the documents indexed with B are a subset of the documents indexed with A. By applying the method, we can find previously unknown taxonomic relations between subject terms without pre-existing information about their relationships. Note that one document can be indexed with multiple subject terms. The subsumption relation between two subject terms, x and y, is identified as follows:

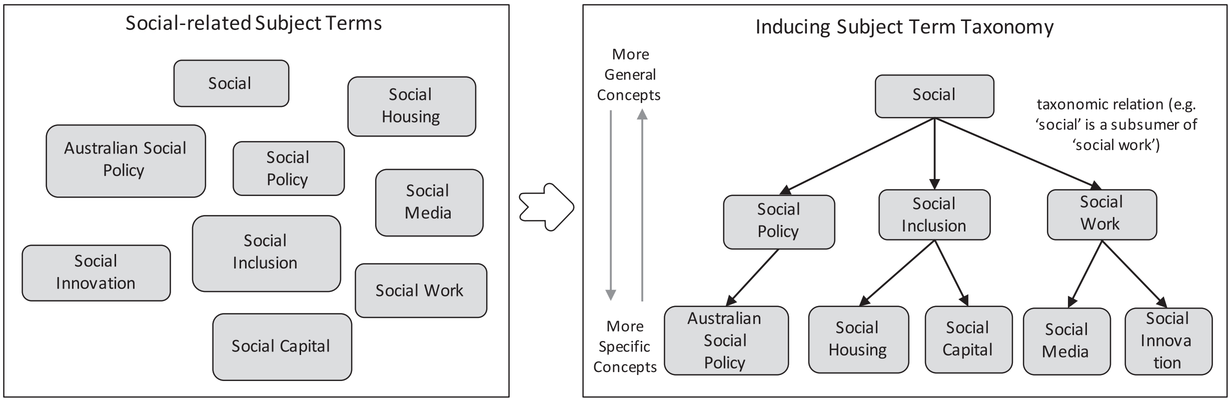


Fig. 3. A taxonomy example consisting of social-related subject terms used in APO.

$$p(x|y) \geq \alpha, \ \ p(y|x) < \alpha, \tag{4}$$

where $p(x|y)$ denotes a conditional probability that x appears, given that y appears for indexing the documents in $\mathscr{D}$, and α is a co-occurrence threshold. This formula is interpreted as follows: x is considered a *subsumer* (*ancestor*) of y, (1) if x appears for indexing more than α proportion over the documents that y appears for indexing and (2) if y appears for indexing less than α proportion over the documents that x appears for indexing. Note that in $\mathscr{S}'$, each subject term is associated with its synonyms (i.e., its secondary subject terms). Thus, in our approach, the occurrences of each subject term $s \in \mathscr{S}'$ is the sum of the occurrences of s and its synonyms.

However, this formula allows a subject term to have its multiple subsumers. This violates the structure of a taxonomy, assuming that one subject term can be subsumed by at most one other subject term. To address this, [11] proposed a subsumption score of a subsumer a for a given subject term x. The score is then used to find the unique subsumer of x. The score is denoted as $ss(a,x)$:

$$ss(a,x) = p(a|x) + \sum_{a' \in S_a} w(a',x) \cdot p(a'|x), \tag{5}$$

where a is a candidate subsumer of x, S_a is the set of subsumers of x, and $w(a',x)$ denotes the weight of the relation between a' and x, calculated by $w(a',x) = \frac{1}{d(a',x)}$, where $d(a',x)$ is the layer distance between a' and x. If the distance is higher, then lower weight is given. Thus, if a subject term x has more than two candidate subsumers S_a, we apply Eq. (5) and choose the subsumer with the highest ss score as the best subsumer for x.

The key parameter in the subsumption method is α in Eq. (4). The higher α is, the lower the average depth and the higher the quality of the induced taxonomy. A trade-off thus needs to be considered between a higher average depth and a higher quality of taxonomic relations. In order to determine a good value of α using a method, we use the *harmonic mean* of a quality and the average depth of the induced taxonomy. As the quality metric, [11] used a metric (called *taxonomic F-measure*) assuming that the gold-standard taxonomy exists. However, this metric cannot be applied where there is no such a taxonomy. In our approach, our premise is that a good taxonomy maximises overall semantic similarities between all the pairs of parent-child in the induced taxonomy. Thus, we choose the notion of similarity as the quality metric. This is aligned with the notion of a good taxonomy used in the prior works [35]. As the similarity measure, we use the same approaches discussed in Section 3. Thus, our objective here is to induce a taxonomy $\mathscr{T}$ such that:

$$\mathscr{T} = \underset{\mathscr{T}_k \subseteq G}{arg\ max} \left(\frac{2ab}{a+b} \right), \tag{6}$$

where $\mathscr{T}_k$ denotes the taxonomy built using $\alpha = k$; G denotes the set of n taxonomies with n values of α, that is, $G = \{\mathscr{T}_{\alpha_1}, \ldots, \mathscr{T}_{\alpha_n}\}$, where α_i is the i-th value of α in G. In this work, we examine values from 0.1 to 0.9 with a step of 0.1 as α values, thus n is set to be 9. The symbol a is the average similarities of all the parent-child pairs of the nodes in $\mathscr{T}_k$, defined as:

$$a = \frac{1}{|\mathscr{T}_k| - 1} \sum_{(x,y) \in \mathscr{T}_k} sim(x,y), \tag{7}$$

for all parent-child pairs of $(x,y) \in \mathscr{T}_k$; $sim(x,y)$ is the similarity function, and $|\mathscr{T}_k|$ is the number of nodes in $\mathscr{T}$. Note that in each taxonomy $\mathscr{T}$ that we built, the number of edges is $|\mathscr{T}| - 1$. The symbol b is the average depth of $\mathscr{T}_k$ denoted as

$$b = \frac{1}{|\mathscr{T}_k|} \sum_{v \in \mathscr{T}_k} depth(v), \tag{8}$$

where $depth(v)$ is the depth of a node v that is the number of edges from v to root node. The root is at depth zero.

5. A case study: analysis of the APO repository

We conduct a case study, where we apply the proposed methodology to the APO repository. First, we analyse how Step 1 in Fig. 1 can identify missing subject terms. Second, we qualitatively measure the quality of the refined subject terms produced by Step 2 in Fig. 1 based on the APO curators. Finally, we present the result of a subject term taxonomy induced from the refined subject terms using Step 3 in Fig. 1. In Section 6, we further present an experiment to evaluate the effectiveness of our taxonomy inducing method using the gold-standard MeSH taxonomy to show its generalisability.[13] The entire APO repository was collected from APO in Jun 2019, where it contained a total of 40,533 documents $\mathscr{D}$ and a total of the existing 5725 subject terms $\mathscr{S}$ used to index $\mathscr{D}$. The assignment of the subject terms to each document $d \in \mathscr{D}$ has been achieved manually by a number of the APO curators based on their understanding, knowledge and experience. Each document consists of its title and text description, and also d is associated with a number of the assigned subject terms.

5.1. Analysis of identifying missing subject terms

Given the subject terms $\mathscr{S}$, we applied our string matching method in Section 3.1 to $\mathscr{D}$ to identify missing subject terms $\mathscr{S}_m$. Fig. 4(a) and (b) show the top-20 most frequently used subject terms before/after identifying $\mathscr{S}_m$ for indexing $\mathscr{D}$. The x-axis represents the document frequencies (DFs) of the subject terms, while the y-axis shows the top-20 subject terms in terms of their DFs. Here, the DF of a subject term s means the number of documents that have s as an indexed subject term. Before identifying $\mathscr{S}_m$, 'regional australia institute in[form] library' is the most dominant subject term in terms of DF, while 'government' is the most dominant after identifying $\mathscr{S}_m$. As observed, overall, the top-20 subject terms are easily distinguished between before/after identifying $\mathscr{S}_m$ and also their DFs are largely different.

From another angle, the overall number of subject terms assigned to each document is notably increased after identifying $\mathscr{S}_m$. We found that the number of the unique subject terms in $\mathscr{D}$ before and after identifying $\mathscr{S}_m$ is 5654 and 5702, respectively. Fig. 5 shows the comparison between two distributions of subject terms used to index $\mathscr{D}$. The blue/green line shows the distribution of the numbers of the assigned subject terms over each document in $\mathscr{D}$ before/after identifying $\mathscr{S}_m$. The x-axis denotes the sorted document indices in $\mathscr{D}$ in terms of such numbers. From this figure, we observe the following: (1) the average number of the subject terms used to index $\mathscr{D}$ before and after identifying $\mathscr{S}_m$ is 4 and 15, respectively; and (2) the maximum number of the subject term used to index $\mathscr{D}$ before and after identifying $\mathscr{S}_m$ is 27 and 134, respectively. As a result, we can find additional subject terms by scanning the text description of $\mathscr{D}$. The subject term set $\mathscr{S}$ now is enlarged and additionally includes the identified missing subject terms.

5.2. Filtering out irrelevant subject terms

Our method for filtering out irrelevant subject terms has been presented in Section 3.2. First, we divide the subject terms $\mathscr{S}$ into the secondary and primary subject terms, denoted as $\mathscr{S}_s$ and $\mathscr{S}_p$, respectively. Thus, $\mathscr{S}_s \cup \mathscr{S}_p = \mathscr{S}$. As discussed in Section 3.2, for this, we calculate the DF of each subject term $s \in \mathscr{S}$. If the DF $\geq$ `min-df`, we assign s to $\mathscr{S}_p$, and $\mathscr{S}_s$ otherwise. To determine a value for `min-df`, we find the distribution of the DF values of the subject terms $\mathscr{S}$ as seen in Fig. 2. As observed, the majority of the subject terms are very infrequently used for indexing, as evident by the long tail of their low DFs. We choose a value for `min-df` capable of selecting only the top 20% of subject terms from $\mathscr{S}$ in terms of their DF values. Thus, a value of 78 is

[13] We also encourage the user to refer to '8' to see examples of the outcomes of the proposed methodology after reading this section.

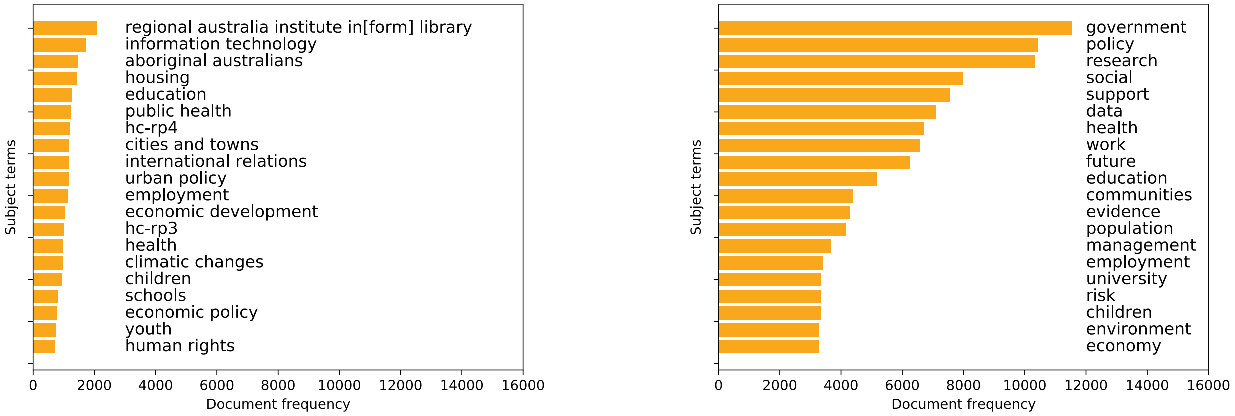


Fig. 4. Top-20 subject terms (y-axis) in terms of their document frequencies for indexing $\mathscr{D}$: the left (right) shows such document frequencies before (after) identifying missed subject terms.

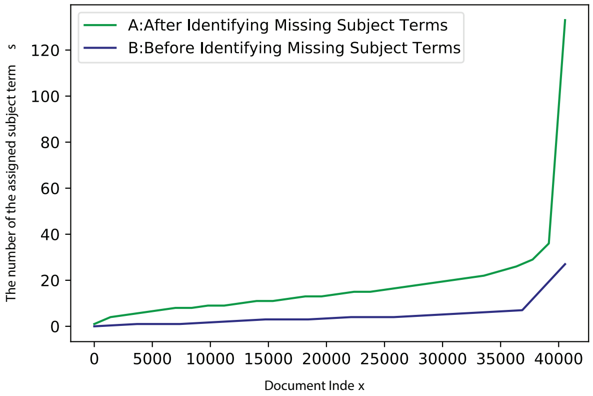


Fig. 5. The distributions of subject terms for indexing the document collection $\mathscr{D}$.

Table 1
The examples of the 10 secondary subject terms.

No.	Secondary subject terms $\mathscr{S}_s$	DF
1	motor vehicles	75
2	seniors	70
3	vegetables	66
4	guarantees	62
5	happiness	56
6	coding	48
7	poisoning	47
8	restaurants	44
9	hygiene	38
10	professional associations	28

chosen as `min-df` for generating $\mathscr{S}_s$ and $\mathscr{S}_p$. As a result, from the total 5702 subject terms in $\mathscr{S}$, we choose the 1146 ones as S_p whose DF values are $\geq$ 78, and the other 4556 subject terms as S_s. Table 1 shows the examples of the 10 subject terms in $\mathscr{S}_s$ with their DFs.

Second, we measure the usefulness of each subject term $s \in \mathscr{S}_s$. If we find $s^{'} \in \mathscr{S}_p$ whose similarity with s is greater than or equal to `min-sim-syns`, we add s to the synonym list of s'. Otherwise, we ignore s by removing s from $\mathscr{S}$. The similarity metric has been discussed in Section 3. To build a word embedding model from $\mathscr{D}$, after removing stopwords and applying lemmatization from $\mathscr{D}$, we used the gensim Word2vec function[14] with the parameters: size (dimensionality of the word vectors): 100, window (maximum distance between the current and predicted word within a sentence): 5, min_count (ignores all words with total frequency lower than this): 3. As a value of `min-sim-syns`, we have chosen 0.7 in our case study.

To assess the effectiveness of our refining methodology for filtering out irrelevant subject terms, we conducted an evaluation based on human judgement using two APO curators. Given the 4556 subject terms in $\mathscr{S}_s$, we further reduced the number of subject terms in $\mathscr{S}_s$ that were to be assessed for their usefulness. As such candidates, we only considered the subject terms in $\mathscr{S}_s$ whose similarity scores with any $s^{'} \in \mathscr{S}_p$ are over 0.7. We assume that if there is no $s^{'} \in \mathscr{S}_p$ whose similarity score is $\geq$0.7 with $s \in \mathscr{S}_s$, we regarded s to be irrelevant, thus being ignored. By doing so, we finally chose the 3008 subject terms and the relevance of each one of them was assessed by the APO curators.

Specifically, given each s of the 3008 terms with its best similar one $s^{'} \in \mathscr{S}_p$ measured by the similarity metrics in Section 3, the APO curators were asked to assess whether s can be correctly assigned as a synonym of s'. If relevant, 1 is given, and otherwise 0 (irrelevant). These curators provided their agreed judgement about ‘relevant’ and ‘irrelevant’. We measure the accuracy of this assessment by dividing the number of values of 1 (relevant) by 3008. The accuracy turns out to be 86.71%. Although there is no universal standard for assessing the result, this result may show that our method for filtering out irrelevant subject terms can be useful and generate a more precise set of dominantly used subject terms for indexing $\mathscr{D}$. Also, the result indicates that our method can contribute to avoiding the unnecessary increase of the vocabulary size of subject terms and helping human indexers (also machines) to choose more relevant subject terms for indexing a new document. Further, by associating secondary subject terms with primary subject terms as synonyms, we can help human indexers to minimise the misinterpretation or ambiguity of the meanings of subject terms.

From another angle, we are interested in analysing accuracy as we increase `min-sim-syns`. We assume that the higher it is, the better accuracy we would expect to get. Thus, we analysed the assessment results based on different values of `min-sim-syns`, {0.7, 0.8, 0.9}. The result can be seen in Table 2. As observed, our assumption is proven to be correct: accuracy gradually increases as we increase `min-sim-syns`. We suggest that an optimal value of `min-sim-syns` could be decided based on more empirical studies or with the help of the curators in the target digital repository.

Table 3 shows the examples of 10 secondary subject terms associated

[14] https://radimrehurek.com/gensim/models/word2vec.html

Table 2
The assessment result of filtering out secondary subject terms.

min-sim-syns	Accuracy (%)
0.7	86.71
0.8	90.46
0.9	93.86

Table 3
The 10 Examples of secondary subject terms associated with the primary subject terms with their similarities.

No.	Secondary subject terms	Primary subject terms	similarity score
1	downloading of data	data	0.92
2	urban revitalisation	urban development	0.88
3	student learning	students	0.86
4	employment policy	employment	0.86
5	social wellbeing	well-being	0.86
6	housing and health	housing	0.85
7	information literacy	literacy	0.85
8	indigenous education	aboriginal Australians education	0.84
9	popular music	music	0.83
10	urban water	water	0.81

with some primary subject terms as synonyms in our case study. Also, the similarity score of each pair is observed in the table. Intuitively, the associated secondary subject terms can be observed closely relevant to the corresponding primary subject terms in terms of their semantics.

5.3. Inducing subject term taxonomy

In this section, we present the result of the induced subject term taxonomy from the refined subject terms $\mathscr{S}'$ using the subsumption method (SS). To show its relative capability, we also compare it with TaxoFinder [5]. TaxoFinder builds a CGraph (Concept Graph) that represents how subject terms are associated together based on their co-occurrences for indexing each document. In the CGraph, nodes are subject terms and an edge represents the co-occurrence of the two nodes. The associative strength of an edge is measured based on the similarity of the two nodes. As the similarity, we used a well-known method, Pointwise Mutual Information (PMI). The notion of PMI is to estimate the likelihood of co-occurrence of two terms, considering their individual frequencies. In our experiment, we rather use the normalised PMI function that returns values between −1 and 1. Since our similarity score range is [0,1], we simply convert the negative PMI values into 0. The normalised PMI function, denoted as *npmi*, for a pair of subject terms *x* and *y* is given as:

$$pmi(xy) = log\frac{p(xy)}{p(x)p(y)},\ npmi(xy) = \frac{pmi(xy)}{log(p(xy))}, \tag{9}$$

where $p(x)$ (resp. $p(y)$) represents the number of occurrences of x (resp. y) over the total occurrences of subject terms in $\mathscr{S}'$, and $p(x,y)$ is the co-occurrences of x and y over the total occurrences of all the pairs of subject terms in $\mathscr{S}'$. In the CGraph, one node can have multiple edges to connect to other nodes. Given the CGraph, TaxoFinder applies the Maximum Spanning Tree (MST) algorithm to induce a taxonomy. The MST is a subset of the CGraph that includes all of $|\mathscr{S}'|$ nodes with the $(|\mathscr{S}'| - 1)$ edges, maximising the associative strengths between the nodes. To apply the MST algorithm, the root node was chosen as the most frequently occurred subject term, assuming that it is likely to be the most general node as used in TaxoFinder.

We now present how we chose an optimal value for α in Eq. (4) for SS. As discussed, to choose such a value, we measured the harmonic mean of the average similarity between all the parent-child nodes in $\mathscr{T}_i$ and the average depth at each α candidate from $\{0.1,\ldots,0.9\}$. Recall that the notion of the similarity is used to represent the quality of the induced taxonomy $\mathscr{T}_i$. The higher, the better quality $\mathscr{T}_i$ reflects. Fig. 6(a) shows the distributions of these two kinds of measured values (similarities and depths) across the different α candidates. The red line shows the distribution of the average similarities and black line is the distribution of the average depths over the candidates. Also, recall that the higher a value for α is, the lower the average depth and the higher the quality of $\mathscr{T}_i$. From this figure, we calculated the harmonic mean values across the same candidates and chose 0.2 as an optimal value for α. The distribution of the harmonic mean values is depicted in Fig. 6(b).

Fig. 7 shows partial taxonomic relations drawn from $\mathscr{T}_i$. We can observe reasonable broader-narrower relations between the subject terms whose semantics are similar. As seen, 'law' is the most abstract term, and a more abstract term is positioned in a higher level in the taxonomy. Also, 'is-a' relationships are easily drawn, for example, 'copyright law' is 'law' (in other words, 'copyright' is a child of 'law'). Further, we can see that the similarity between 'law' and 'crime' is higher than the similarity between 'law' and 'fraud'.

Fig. 8(a) and (b) show the entire structures of the taxonomy (denoted as $\mathscr{T}_x$) induced using SS and the one (denoted as $\mathscr{T}_y$) induced using TaxoFinder. The root node is indicated by the node in red, while other subject terms are denoted by blue circles. As can be seen, SS generated a more stable and balanced taxonomy in terms of the descendent distribution than $\mathscr{T}_y$. That is, TaxoFinder generated a significantly deeper taxonomy than $\mathscr{T}_x$. It turned out that the average depth of $\mathscr{T}_x$ is 4 and that of $\mathscr{T}_y$ is 34. The max depth of $\mathscr{T}_x$ is 9 and that of $\mathscr{T}_y$ is 57. One possible reason is that TaxoFinder uses a MST algorithm to build a final taxonomy based only on the associative strengths between nodes, and it does not incorporate any method for finding an optimal depth of the induced taxonomy. Also we can see that some nodes have only one child node in $\mathscr{T}_y$. Also, when comparing the top-level terms spreading out horizontally, $\mathscr{T}_x$ shows more broad range of subject terms (i.e., five) than $\mathscr{T}_y$ showing only one subject term. As we see in each taxonomy, subject terms are connected to each other on different depths, and semantic similarities between them are easily drawn by navigating their paths in the taxonomy.

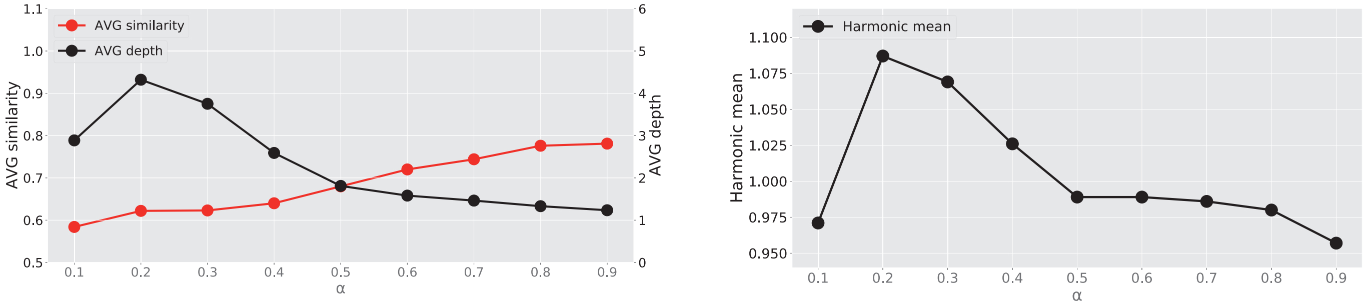
In order to better understand the effectiveness of the subsumption method (i.e., SS), we present the quantitative comparison between SS and TaxoFinder in the next section.

6. Evaluation of inducing subject term taxonomy

We now quantitatively evaluate the effectiveness of the subsumption method (SS) using a publicly well-known dataset. This can improve our understanding of the generalisability of our approach for inducing a taxonomy. To show the relative effectiveness of SS, we also compare it with TaxoFinder. We assess the induced taxonomy $\mathscr{T}_i$ from each method (i.e., SS or TaxoFinder) by comparing it with the existing gold-standard taxonomy $\mathscr{T}_g$. Our aim to induce a taxonomy $\mathscr{T}_i$ as much close as possible to $\mathscr{T}_g$.

6.1. The corpus for MeSH

MeSH was chosen as the source of the gold-standard taxonomy $\mathscr{T}_g$. MeSH is a representative, biomedical controlled vocabulary consisting of 29k+ biomedical subject terms in a taxonomy introduced by National



(a) Distribution of the average similarities vs. depths

(b) Distribution of the harmonic mean values

Fig. 6. The distributions of the harmonic mean values using average similarities and average depths of the induced taxonomy across different α values from the APO data.

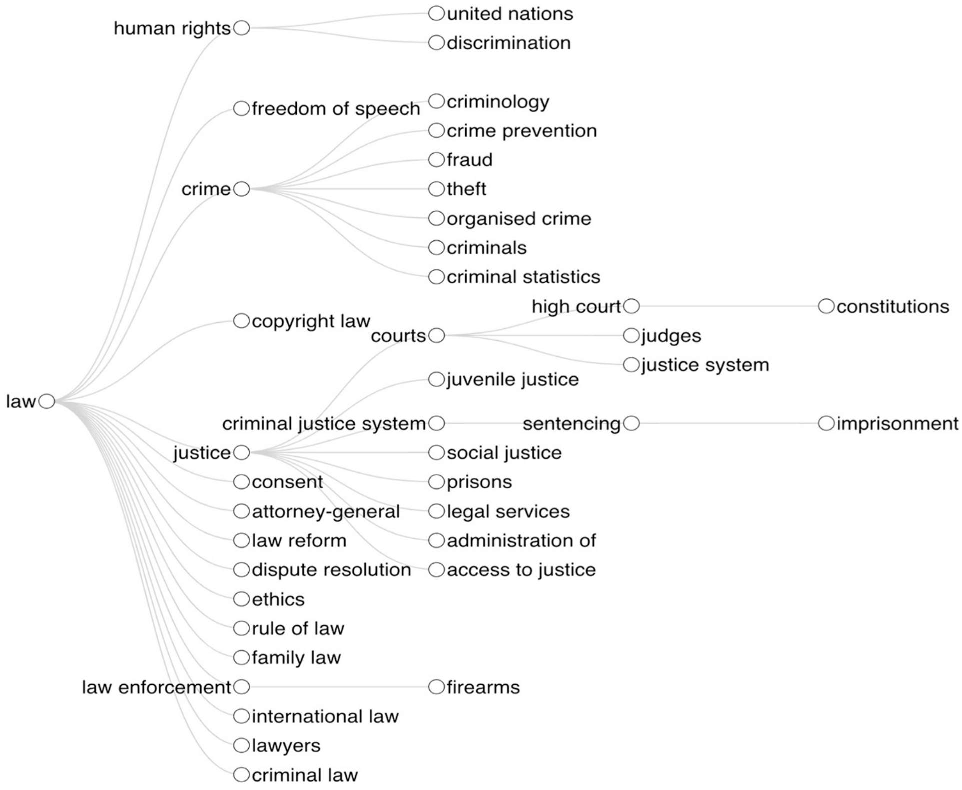


Fig. 7. An example of taxonomic relations in the induced taxonomy from the APO data.

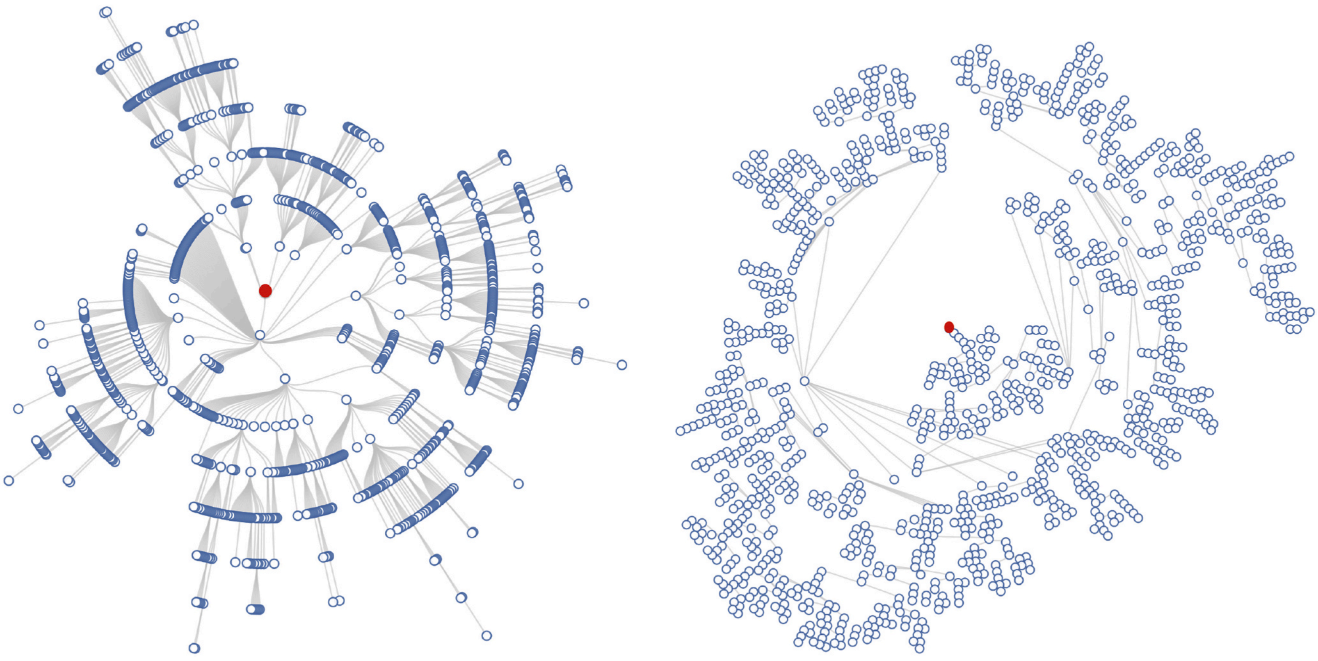
Library of Medicine. Our rationale for choosing MeSH is two-fold: First, it is one of well-defined and publicly accessible tree structures of medical subject terms in biomedical science. Second, BioASQ[15] also publicly provides the 14 millions of annotated biomedical articles (title + abstract) by MeSH terms, where ‘annotated’ means that MeSH terms have been assigned to the articles in MEDLINE® as indexing terms. MeSH terms are assigned to each article in MEDLINE®, in order to describe what the article is about. Thus, our hypothesis is that by inducing a MeSH term taxonomy from such articles using SS and comparing it with the gold-standard MeSH, we aim to achieve our generalisability study of SS.

Further, we aimed to conduct a more precise, simpler validation using MeSH. For this, we arbitrarily chose one of the 16 main branches in MeSH. We selected the ‘diseases’ taxonomy with the biomedical articles indexed with the descendent subject terms of the MeSH subject term ‘diseases’. Thus, the root node of T_g is ‘diseases’. This taxonomy $\mathscr{T}_g$ has the largest number of immediate children (i.e., 26) among the 16 branches of MeSH. It has the depth of 10, and the number of its descendent MeSH terms is 10k+ which is the second largest number of descendants while the ‘Chemist and Drugs’ branch has the largest number of descendant MeSH terms (i.e., 19k+).

In 2020, we downloaded the 2020 version[16] of MeSH and the biomedical articles indexed by the MeSH terms from BioASQ. From these articles, we only chose the articles which have been published in the last 5 years and also indexed by the terms under the *diseases* subtaxonomy. Then, we obtained around 630,336 articles. Afterwards, we ignored relatively lower document frequencies (DFs) of MeSH terms to reduce a bias of DFs of the MeSH terms. For this, we ignored the 25% of

[15] http://participants-area.bioasq.org/datasets/

[16] https://www.nlm.nih.gov/databases/download/mesh.html



(a) Using the subsumption method

(b) Using TaxoFinder

Fig. 8. Induced taxonomies.

the MeSH terms with the lowest DFs. The total number of the MeSH terms used in our experiment turned out to be 836. Further, from the 630,336 articles, we filtered out certain articles whose the number of indexed MeSH terms under the 836 *diseases* subtaxonomy are < 2, as these articles do not have co-occurred MeSH terms. Note that the subsumption method is based on co-occurrence of subject terms. Finally, the 82,799 articles with the 836 MeSH terms were used as the input to SS. The 630,336 articles are used to build a word embedding model using Word2vec [33]. To build the embedding model, we used the same parameters using the gensim Word2vec method as done in Section 5.2. Also, note that we reorganised the *diseases* subtaxonomy using only the 836 terms $\mathscr{S}$, maintaining their ancestor-descendent relations. For example, given a term $s \in \mathscr{T}_g$, if s's parent p is not included in the 836 terms, we connected s with the p's immediate ancestor that is included in $\mathscr{S}$.

6.2. Evaluation metrics

As the evaluation metrics, we use *global taxonomic F-measure* (TF), the harmonic mean of *global taxonomic precision* (TP) and *global taxonomic recall* (TR), as widely used for assessing induced taxonomies [5]. Calculating these metrics can be divided into a *local* and a *global* measure. The local measure is used for comparing the position of a term in the induced taxonomy $\mathscr{T}_i$ with the position of the same term in T_g. The global measure then aggregates the local measures of all terms in $\mathscr{T}_i$ and calculates the overall taxonomic quality of $\mathscr{T}_i$. As the local measure, the notion of the *common semantic cotopy* (csc) is used [5]. Given a term s, $\mathscr{T}_i$ and $\mathscr{T}_g$, their csc denotes the set that includes s and its ancestors and descendants shared by both $\mathscr{T}_i$ and $\mathscr{T}_g$. Formally, the csc is defined as:

$$csc\left(s,\mathscr{T}_i,\mathscr{T}_g\right) = \left\{s_i | s_i \in S_i \cap S_g \ \left(s_i \le_{S_i} s \vee s \le_{S_i} s_i\right)\right\}, \tag{10}$$

where S_i (resp. S_g) is the set of terms in $\mathscr{T}_i$ (resp. $\mathscr{T}_g$) (in our work, $S_i = S_g$), and '$\le_{S_i}$' is the order induced by taxonomic relations in $\mathscr{T}_i$ (i.e., s_i is either an s's descendent ($s_i < s$) or an ancestor ($s < s_i$) or $s_i = s$). Using this notion, *local taxonomic precision (tp)* and *local taxonomic recall (tr)* are defined as:

$$\begin{aligned} tp(s) &= \frac{\left|csc\left(s,\mathscr{T}_i,\mathscr{T}_g\right) \cap csc\left(s,\mathscr{T}_g,\mathscr{T}_i\right)\right|}{\left|csc\left(s,\mathscr{T}_i,\mathscr{T}_g\right)\right|}, \\ tr(s) &= \frac{\left|csc\left(s,\mathscr{T}_i,\mathscr{T}_g\right) \cap csc\left(s,\mathscr{T}_g,\mathscr{T}_i\right)\right|}{\left|csc\left(s,\mathscr{T}_g,\mathscr{T}_i\right)\right|}. \end{aligned} \tag{11}$$

Then, TP, TR, and TF are finally drawn from the local estimation:

$$\begin{aligned} \mathrm{TP} &= \frac{1}{\left|S_i \cap S_g\right|} \sum_{s \in S_i \cap S_g} tp(s), \\ \mathrm{TR} &= \frac{1}{\left|S_i \cap S_g\right|} \sum_{s \in S_i \cap S_g} tr(s), \\ \mathrm{TF} &= \frac{2 \cdot \mathrm{TP} \cdot \mathrm{TR}}{\mathrm{TP} + \mathrm{TR}}, \end{aligned} \tag{12}$$

where the higher a TF is, the better quality of $\mathscr{T}_i$ is considered.

We now analyse the evaluation result using the MeSH data. First, as an optimal value α for SS, 0.1 was chosen using the same method as done with the APO data. Similar to the result using the APO data, it turned out that SS generated a better quality of taxonomy, producing the average and max depths are 3 and 40, respectively, while TaxoFinder produced 40 and 75 respectively. Thus, TaxoFinder produced the taxonomy with a much greater depth. Our focus is now to present the quality of the induced taxonomy using SS, in comparison with TaxoFinder, by means of TP, TR, and TF. Table 4 shows the comparison of these two methods, where the higher score under each metric is denoted in bold. As observed, SS outperforms TaxoFinder in TP and TF. In particular, in TF, SS turns out to be more than 60 times better than TaxoFinder. Although

Table 4
The comparison between the subsumption method (SS) and TaxoFinder in terms of TP, TR, and TF.

Method	TP	TR	TF
SS	**0.524**	0.831	**0.642**
TaxoFinder	0.051	**0.860**	0.097

TaxoFinder shows a slightly better performance on TR, the overall performance (i.e., TF) is much lower than SS. Thus, our evaluation results show that our induced taxonomy shows a much higher quality, in comparison with the state-of-the-art method, TaxoFinder. The TF score 0.642 roughly indicates that 64% of the taxonomic relations, identified by SS, are the same with the relations in the gold-standard *diseases* taxonomy. This TF result is almost similar to the highest TF results of the proposed approaches in related works using different datasets [5,11]. Finally, our evaluation results show the validity that SS has a potential to be used for identifying semantics and broader-narrower relations of subject terms.

7. Conclusion

In this paper, we presented a methodology for refining an existing set of subject terms used to index a digital collection in a digital repository. The motivation of our work was to (1) additionally find potentially relevant subject terms that have been missed for indexing the collection, and (2) also generate a more precise, meaningful set of subject terms by refining the existing subject terms. Further, we presented the method for inducing a taxonomy from the refined subject terms, with our proposed objective function that maximises the harmonic mean of the quality of the induced taxonomy $\mathscr{T}_i$ using the notion of similarity between the nodes in $\mathscr{T}_i$ and its depth. We evaluated the quality of the refined subject terms based on human judgement using the APO repository. Also, to show the generalisability of our taxonomy inducing method, we quantitatively measured its effectiveness using the MeSH taxonomy with the document collection downloaded in the year of 2020 from BioASQ, in comparison with the state-of-the-art taxonomy learning method, TaxoFinder [5]. Our evaluation showed that the proposed methodology has high potential for refining an existing set of subject terms and capturing their semantic relationships. Our methodology has been designed to be generalisable, thus can be applied to various repositories or domains, where a document collection and its indexed subject terms exist. We expect that this study can provide two major benefits for the digital library community: (1) improving the ability to index the underlying document collection with more precise, relevant subject terms; and (2) providing better understanding about the semantic relationships among underlying subject terms from an induced taxonomy, helping to produce more accurate indexing of the collection.

There may be some possible limitations in this study. To identify synonyms, this study proposed the similarity measures using WordNet and the word embeddings. However, we may need more thorough experiments to validate the effectiveness of these similarity measures in comparison with various similarity measures. Also, this work does not address how to incrementally evolve the structure of the induced taxonomy as we have more documents and their indexed subject terms. Considering that digital documents are growing rapidly these days, it would be interesting to keep the induced taxonomy up-to-date, evolving and reflecting changes occurred in the current document collection.

As future work, we plan to conduct more comprehensive case studies to further validate the methodology. Also, we will include a testing with the end-users of the APO repository for determining whether their needs can be better met by using the refined subject terms for indexing. Further, we plan to investigate an automatic method for reflecting refining subject terms into the APO metadata records, and improve the document indexing task using the metadata.

CRediT authorship contribution statement

Yong-Bin Kang: Investigation, Conceptualization, Methodology, Software, Formal analysis, Writing - review & editing. **Jihoon Woo:** Software, Formal analysis, Writing - review & editing. **Les Kneebone:** Conceptualization, Methodology, Writing - review & editing. **Timos Sellis:** Project administration, Funding acquisition, Supervision, Investigation, Conceptualization, Writing - review & editing

Acknowledgements

This work was supported by the Linked semantic platforms for social & physical infrastructure & wellbeing project, LE180100094, funded by an Australian Research Council's Linkage Infrastructure, Equipment and Facilities (LIEF) grant. We especially thank Associate Professor Brigid van Wanrooy, Director of Analysis and Policy Observatory (APO), for providing insightful comments and support during the research project.

Appendix A. Demonstration using an article in APO

In this appendix, we demonstrate the results of identifying missing subject terms, refining subject terms, and inducing a subject term taxonomy using a short article that has only an abstract from APO repository.

- **Title**: Improving Indigenous completion rates in mainstream TAFE: an action research.
- **Body Text**: Indigenous engagement with vocational education and training (VET) hasimproved significantly, but successful Indigenous completion rates are lower nationally when compared to the overall population. In this report, based on an action research project, Jo Balatti, Lyn Gargano, Martha Goldman, Gary Wood and Julie Woodlock examine intra-institutional factors at four Queensland TAFE institutes to better understand and take action on issues affecting Indigenous completion rates. At policy level, the authors conclude mainstream programs require examination at three levels - intellectual, cultural and social - to develop effective responses to facilitate successful Indigenous completion rates. At organisation level, the authors recommend examination of organisational culture for consistency of values and beliefs, and practices in terms of content, teaching, support, and collaboration.
- **Existing subject terms**: Aboriginal Australians
- **Identified missing subject terms**: TAFE, VET, indigenous, action research, research, collaboration, content, culture, education, education and training, organisational culture, policy, population, social, support, teaching, training, values, vocational education, vocational education and training.
- **Refined subject terms**: TAFE, VET, indigenous, action research, research, collaboration, content, culture, education, education and training, policy, population, social, support, teaching, training, values, vocational education, vocational education and training. The term 'organisational culture' has been added to the synonym list of 'culture'.
- Fig. A.9 shows the induced taxonomy in our case study in Section 5. The subject terms used in this example are denoted in red, while the subject terms in black were added to help to understand their relationships with the red ones.

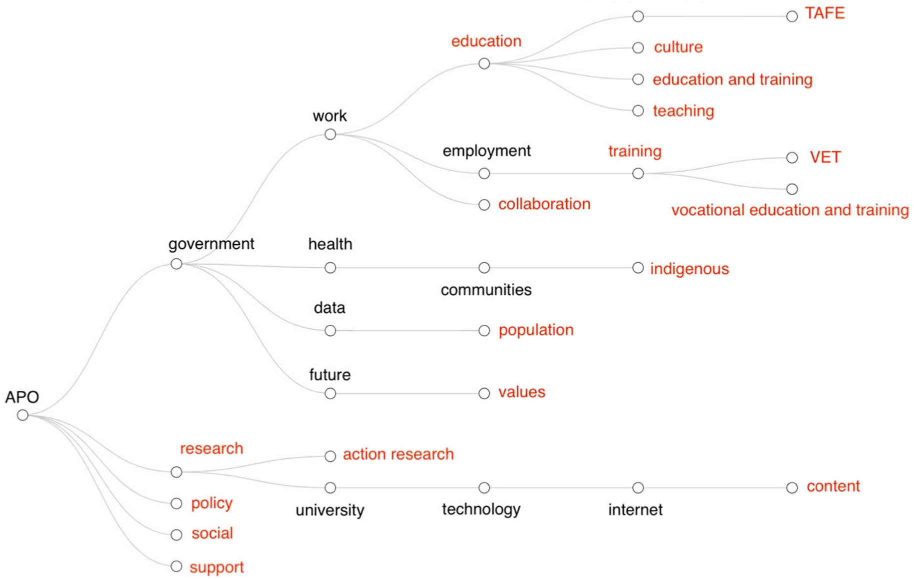


Fig. A.9. Induced subject term taxonomy with refined subject terms from an APO document.

References

[1] I. Xie, K. Matusiak, Discover Digital Libraries: Theory and Practice, 1st ed., Elsevier Science, 2016.

[2] E. Chung, S. Miksa, S.K. Hastings, A framework of automatic subject term assignment for text categorization: an indexing conception-based approach, J. Am. Soc. Inf. Sci. Technol. 61 (4) (2010) 688–699.

[3] L. Kneebone, A subject vocabulary for the social policy and research sector: Leveraging a keyword archive, in: VALA2020, 2020.

[4] H. Tarver, M. Phillips, O. Zavalina, P. Kizhakkethil, An exploratory analysis of subject metadata in the digital public library of america, in: Proceedings of the 2015 International Conference on Dublin Core and Metadata Applications, Dublin Core Metadata Initiative, 2015, pp. 30–40.

[5] Y. Kang, P.D. Haghigh, F. Burstein, TaxoFinder: a graph-based approach for taxonomy learning, IEEE Trans. Knowl. Data Eng. 28 (2) (2016) 524–536.

[6] D. Haynes, Metadata for Information Management and Retrieval: Understanding Metadata and its Use, Facet Publishing, 2018, 2nd Revised Edition (May 22, 2018).

[7] R.F. Smallwood, Information Governance: Concepts, Strategies and Best Practices, 2nd edition, Willey, 2019.

[8] J. Greenberg, Metadata capital: raising awareness, exploring a new concept, Bull. Assoc. Inf. Sci. Technol. 40 (4) (2014) 30–33.

[9] R. Bennett, E.T. O'Neill, K.A. Kammerer, Assignfast: an autosuggest based tool for fast subject assignment, Inf. Technol. Libr. 33 (2014) 34–43.

[10] S. Huang, X. Luo, J. Huang, H. Wang, S. Gu, Y. Guo, Improving taxonomic relation learning via incorporating relation descriptions into word embeddings, Concurr. Comput. 32 (14) (2020), e5696.

[11] K. Meijer, F. Frasincar, F. Hogenboom, A semantic approach for extracting domain taxonomies from text, Decis. Support. Syst. 62 (2014) 78–93.

[12] H. Hedden, Taxonomies and controlled vocabularies best practices for metadata, J. Digital Asset Manag. 6 (5) (2010) 279–284.

[13] O. Medelyan, Human-Competitive Automatic Topic Indexing, University of Waikato, 2009. https://books.google.com.au/books?id=7AqinQAACAAJ.

[14] A. Kühnemund, The role of applications within the reviewing service zbmath, PAMM 16 (1) (2016) 961–962.

[15] S. Sunny, M. Angadi, Applications of thesaurus in digital libraries, DESIDOC J. Library Inform. Technol. 37 (5) (2017) 313–319, https://doi.org/10.14429/djlit.37.5.11169.URL. https://publications.drdo.gov.in/ojs/index.php/djlit/article/view/11169.

[16] M. George, M. Emma, Collaborative tagging as a knowledge organisation and resource discovery tool, Libr. Rev. 55 (5) (2006) 291–300.

[17] H. White, C. Willis, J. Greenberg, HIVEing: the effect of a semantic web technology on inter-indexer consistency, J. Doc. 70 (2014) 307–329.

[18] J. Greenberg, Metadata generation: processes, people and tools, bulletin of the American Society for information science and technology, Bull. Am. Soc. Inf. Sci. Technol. 29 (2) (2005) 16–19.

[19] R.C. Knight, E. Rodrigues, R. Ciota, Facilitating collaborative metadata creation for faculty-initiated digital projects, J. Libr. Metadata 20 (1) (2020) 51–64.

[20] K.W. Broman, K.H. Woo, Data organization in spreadsheets, Am. Stat. 72 (1) (2018) 2–10.

[21] Y.-B. Kang, P. Delir Haghighi, F. Burstein, CFinder: an intelligent key concept finder from text for ontology development, Expert Syst. Appl. 41 (9) (2014) 4494–4504.

[22] R. Campos, V. Mangaravite, A. Pasquali, A. Jorge, C. Nunes, A. Jatowt, YAKE! Keyword extraction from single documents using multiple local features, Inf. Sci. 509 (2020) 257–289.

[23] M.A. Hearst, Automatic acquisition of hyponyms from large text corpora, in: Proceedings of the 14th Conference on Computational Linguistics 2, 1992, pp. 539–545.

[24] S.P. Ponzetto, M. Strube, Taxonomy induction based on a collaboratively built knowledge repository, Artif. Intell. 175 (9–10) (2011) 1737–1756, https://doi.org/10.1016/j.artint.2011.01.003.

[25] Y. Song, S. Liu, X. Liu, H. Wang, Automatic taxonomy construction from keywords via scalable bayesian rose trees, IEEE Trans. Knowl. Data Eng. 27 (7) (2015) 1861–1874.

[26] W. Wong, W. Liu, M. Bennamoun, Ontology Learning from Text: A Look Back and Into the Future 44, ACM Computing Surveys, 2012, 4.

[27] B.S. Everitt, S. Landau, M. Leese, Cluster Analysis, 4th ed., Wiley Publishing, 2009.

[28] A. Lawrence, Digital curation of public policy resources: Discovery, access and management for policy and practice, in: VALA2016, 2016.

[29] C. Pappas, I. Williams, Grey literature: its emerging importance, J. Hosp. Librariansh. 11 (3) (2011) 228–234, https://doi.org/10.1080/15323269.2011.587100.

[30] J. Mixter, E.R. Childress, Fast (faceted application of subject terminology) users: summary and case studies, OCLC Research (2013).

[31] E.R. Childress, D. Vizine-Goetz, Faceted application of subject terminology (fast), in: Encyclopedia of Library and Information Sciences, CRC Press, 2017, pp. 1539–1548.

[32] J.J. Jiang, D.W. Conrath, Semantic similarity based on corpus statistics and lexical taxonomy, in: Proceedings of the 10th Research on Computational Linguistics International Conference, 1997, pp. 19–33.

[33] T. Mikolov, K. Chen, G. Corrado, J. Dean, Efficient Estimation of Word Representations in Vector Space, 2013.

[34] I. Yamada, A. Asai, H. Shindo, H. Takeda, Y. Takefuji, Wikipedia2vec: An Optimized Tool for Learning Embeddings of Words and Entities from Wikipedia, 2018.

[35] H. Yang, Constructing task-specific taxonomies for document collection browsing, in: Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning, 2012, pp. 1278–1289.

Yong-Bin Kang received a PhD in Information Technology from Monash University in 2011. Currently, he is a senior data science research fellow for the ARC Centre of Excellence for Automated Decision Making and Society. His research expertise and interests are mainly in the fields of natural language processing (NLP), machine learning, and data mining. He has experience in working, managing and delivering large industrial, multidisciplinary research projects in data science such as patent analytics, clinical data analytics, scientific-article analytics, social-media data analytics, expert finding and matching, and machine learning algorithms and applications. His research has been demonstrated by publications in Information Systems journal, WIREs data mining and knowledge discovery, JMIR public health and surveillance, IEEE Transactions on Knowledge and Data

Engineering, IEEE Transactions on Cybernetics. Some representative conference papers A/A* include AAAI, ECAI, ISWC, CIKM, and K-CAP.

Jihoon Woo received a bachelor's degree in data science from Swinburne University of Technology. He is currently working at Social Data Analytics Lab in Swinburne University of Technology as a research assistant. His research interests include machine learning, natural language processing, and knowledge graph.

Les Kneebone is an Information Architect in Analysis and Policy Observatory in Australia. He has worked in information management roles in government, school, community and research sectors since 2002. He mainly contributed to managing metadata, taxonomies and cataloguing standards used in these sectors. Before graduating in Information Management at RMIT, He gained post-graduate qualifications in Sociology and the History and Philosophy of Science (Queensland University of Technology and and University of Melbourne).

Timos Sellis is a Research Scientist at Facebook (USA) and an Adjunct Professor at Swinburne University of Technology (Australia), where between 2016 and 2020 he was a Professor and the Director of the Data Science Research Institute. He got his Diploma degree in Electrical Engineering in 1982 from the National Technical University of Athens (NTUA), Greece. In 1983 he got the M.Sc. degree from Harvard University (USA) and in 1986 the Ph.D. degree from the University of California at Berkeley (USA), both in Computer Science. He has been in the past a faculty member at the University of Maryland (USA, 1986-92), the National Technical University of Athens (Greece, 1992-2013), and RMIT University (Australia, 2013-2016). In 2018, he was awarded the IEEE TCDE Impact Award, in recognition of his impact in the field and for contributions to database systems research and broadening the reach of data engineering research. His research interests include big data, data streams, personalization, data integration, and spatio-temporal database systems . He is a fellow of the IEEE and ACM.