Mapping Semantic Knowledge for Unsupervised Text Categorisation

Xiohui Tao[1] **Yuefeng Li**[2] **Ji Zhang**[1] **Jianming Yong**[3]

[1] Centre for Systems Biology, University of Southern Queensland, Australia,
Email: {xtao, ji.zhang}@usq.edu.au

[2] Science and Engineering Faculty, Queensland University of Technology, Australia
Email: y2.li@qut.edu.au

[3] School of Information Systems, University of Southern Queensland, Australia,
Email: jianming.yong@usq.edu.au

Abstract

Text categorisation is challenging, due to the complex structure with heterogeneous, changing topics in documents. The performance of text categorisation relies on the quality of samples, effectiveness of document features, and the topic coverage of categories, depending on the employing strategies; supervised or unsupervised; single labelled or multi-labelled. Attempting to deal with these reliability issues in text categorisation, we propose an unsupervised multi-labelled text categorisation approach that maps the local knowledge in documents to global knowledge in a world ontology to optimise categorisation result. The conceptual framework of the approach consists of three modules; pattern mining for feature extraction; feature-subject mapping for categorisation; concept generalisation for optimised categorisation. The approach has been promisingly evaluated by compared with typical text categorisation methods, based on the ground truth encoded by human experts.

Keywords: Text Categorisation, Knowledge Mapping, Ontology

1 Introduction

Aiming to promote information accessibility and data analytic capability, machine learning techniques have been deeply studied and extensively used to solve problems in areas such as information gathering (Tao et al. 2011, 2008), information security (Sun et al. 2011, 2012), knowledge discovery (Zhong et al. 2012) and sentiment analysis (Tjondronegoro et al. 2011), etc. Text categorisation aims to categorise a stream of documents into a set of categories. The semantically meaningful categories can help users to access to the documents in demand even in a large collection. The efficient access leads to further development in many disciplines, such as digital libraries, text mining, and knowledge engineering.

Text categorisation could be with supervised or unsupervised strategy. With a set of samples labelled with accurate categories, the supervised strategy categorises an incoming stream of documents by finding their common features with the samples. The performance relies on (i) the accuracy of categories assigned to the samples. The common features shared by the samples and target documents would lead to only the wrong direction if the labelled categories on samples are wrong; (ii) the effectiveness of feature extraction algorithms. If a poor algorithm was employed, non-descriptive and non-discriminative features would be extracted and map the documents with wrong samples. On the other hand, an effective algorithm might extract descriptive and discriminative features and help in categorisation to reduce noise, which is usually caused by sense ambiguities, sparsity, and high dimensionality of the documents (Hu et al. 2009); (iii) the topic coverage of categories. An ambiguous category could be assigned to a document if their semantic spaces just overlay on the boundary and no better, accurate category could be found. Sometimes although a set of categories with comprehensive topic coverage is available, the large number of classes would easily introduce much noise to the results (Gabrilovich & Markovitch 2005). Text categorisation could also be unsupervised when sample documents with labelled categories are in absence. The "cold start" is a real world sample of the problem. Recommender systems cannot make adequate suggestion to user tagging if no (or not a sufficient number of) good sample tags are seen on the product. Recommender systems employing unsupervised strategy then learn an annotation model from the information associated with the product and recommend the annotated features to the user as the possible tags (categories). Unsupervised categorisation strategy attempts to address the sample absence problem by analysing local features extracted from documents. With unsupervised strategy, the feature extraction issue and category coverage issue would still make large impact to text categorisation performance, though samples are no longer in use. Unsupervised categorisation strategy is highly useful, as in the real world qualified samples cannot be easily guaranteed. However, unsupervised strategy is more challenging. The accuracy of categorisation becomes hard to assure when no experience can be borrowed from existing samples.

Text categorisation could handle single-label or multi-label problems. Traditionally, text categorisation techniques employ the Boolean model yo assign only a single category to a document. A candidate category would be considered either 'no' then left away or 'yes' then assigned to the document. More advanced, the index model is employed for single-label categorisation. The top indexed category would be chosen and assigned to the document. The indexing model is superior to the Boolean model. A category would be compared with not only document features but also other category candidates. However, when choosing only the top candidate, others are neglected, no matter how close they are to the top candidate in term of either indexing value or semantic meaning. If a set of equally (or almost equally) qualified categories is encountered, the single label techniques would force the applications to choose only one of them. As a result, only partial information in the document is recognised. Aiming at addressing the drawback, the multi-label techniques are developed. Still employing the index model, the multi-label techniques incorporate a threshold to draw a borderline and choose a set of category candidates that fall into the qualified space. The multi-label strategy is more natural, because it recognises multi-facets of a document. The multi-label strategy is also highly applicable. In many circumstances multi-labels are required for example, categorising library catalogues into multiple subject headings, especially when dealing with a very large volume of collection (Yang et al. 2009). However, the multi-label strategy has introduced another challenging problem - how to choose a qualified threshold to balance the information gain and information lost in categorisation. A restricted threshold might push away some potentially qualified categories; a relaxing threshold might include noisy categories. Thus, technically the performance of multi-label categorisation is largely affected by the evaluation strategy employed to choose the right threshold.

In this paper, we propose an approach to categorise documents into multiple categories without the requirement of training samples. In order to assure the categorisation result, the approach first analyses a document locally and extracts features by using pattern mining techniques, aiming to deal with the effectiveness issue of feature extraction. The approach also incorporates a world ontology containing a large volume of subjects with extensive coverage of topics. Taking the advantage, the proposed approach uses the subjects as the category base to address the topic coverage issue. A set of potentially qualified subjects are chosen from the ontology, by a method introduced to map the document features (local knowledge) to the ontological subjects (global knowledge). Finally, the approach investigates the semantic relations and ontological structure to generalise the mapping knowledge for optimised category assignment. The knowledge generalisation aims at balancing the information gain and loss in multi-label categorisation. The approach thus, consists of three modules; pattern mining for feature extraction; feature-subject mapping for categorisation; concept generalisation for optimised categorisation. The world ontology is encoded from the Library of Congress Subject Headings (LCSH), which represents the natural growth and distribution of human intellectual work (Chan 2005). The proposed approach has been experimentally evaluated using a large library catalogue, by compared with typical categorisation approaches. The presented work makes following contributions:

- An unsupervised knowledge mapping approach that assigns documents into multiple categories;
- A knowledge generalisation method that takes into account the semantic relations and structure of concepts;
- An exploration of exploiting the Library of Congress Subject Heading to promote accessibility to the textual world;
- An extensive empirical experiment that evaluated the knowledge mapping approach using the real world data.

The paper is organized as follows. Section 2 discusses the related work; Section 3 provides the definitions for the world ontology and the research problem; Section 4 presents the conceptual framework for the design of the proposed method. After that, Section 5 introduces the proposed categorisation approach. The experiment design is described in Section 6, and the results are discussed in Section 7. Finally, Section 8 gives the conclusions and future work.

2 Related Work

Unsupervised text categorisation classifies documents into categories with absence of any pre-labelled samples for training. Without labelled samples, Yang et al. (Yang et al. 2010) have formalised a categorisation model by analysing the correlating auxiliary categories. Similarly, the work presented in the paper does not rely on any pre-labelled samples for training. Instead, it tries to map the local concepts underlying from the document to the global concepts specified in a world ontology, and uses the global concepts as the suggesting categories. The strategy of exploiting global knowledge for text categorisation was also studied by Yan et al. (Yan et al. 2009). In the work, they examined the concept relations in Wikipedia, which is a global ontology with a large volume of concepts, and integrated linguistic analysis with word distribution statistics to improve the performance of semantic relation extraction. Comparing with Yan et al. (Yan et al. 2009), the work in this paper has a different aim of facilitating the development of multi-labelled categorisation, and in practice exploits a different global ontology encoded from the Library of Congress Subject Headings, which is a more formal, human intellectual work under continuous growth and revision in the past century. Cai et al. (Cai et al. 2010) also proposed an unsupervised approach to evaluate and improve the quality of local concepts extracted from documents. The work presented in this paper also incorporates feature extraction for local concepts, and further investigates the local concepts by mapping them with global concepts in an ontology to improve categorisation performance.

Multi-label categorisation has been studied by Yang et al. (Yang et al. 2009), who adopted active learning algorithms for categorisation. Our work, however, exploits a mapping method to bridge local features to global concepts and an algorithm to investigate their semantic relations. Cai et al. (Cai et al. 2010) proposed a Multi-Cluster Feature Selection (MCFS) method for unsupervised feature selection and contributed to image clustering. Our work is also on unsupervised and adopts feature selection strategy. However, the work focuses on text categorisation and exploits a large category base provided by a world ontology.

Ontologies have been exploited by much work to facilitate text categorisation. Gabrilovich and Markovitch (Gabrilovich & Markovitch 2005) attempted to generate features using domain-specific and common-sense knowledge in large ontologies. In comparison, our work moves beyond feature extraction and investigates the concept structure in ontologies. Other work that takes ontological structure into account includes Cheng et al. (Cheng et al. 2011), who investigated structured knowledge to assist feature extraction. Their work is based on the same argument as that in our work; rich knowledge is underlying from the conceptual structure and could be exploited to assist text analysis. Camous et al. (Camous et al. 2007) also counted ontological structure as a valuable source. Using the Medical Subject Headings (MeSH) ontology, they introduced a domain-independent method to observe inter-concept relationships in the ontological structure and categorise documents into the MeSH subjects. Also through domain ontologies, Hernandez et al. (Hernandez et al. 2007) proposed to model context in documents by means of domain ontologies, aiming to better access user information needs. These work focuses on specific domains. As a result, the work may improve the precision of categorisation but has also limited the extensibility to other domains. In comparison, our work uses the LCSH, a world knowledge ontology that has extensive coverage of topics, and has no limit in any specific domains. Another world knowledge ontology, Wikipedia, is employed by Wang and Domeniconi (Wang & Domeniconi 2008) and Hu et al. (Hu et al. 2009). Independently, they derived background knowledge from Wikipedia to represent documents and attempted to deal with the sparsity and high dimensionality problems in text categorisation. Also adopting on Wikipedia, Kiran et al. (Kiran et al. 2010) attempted to cluster documents by finding a score based on the analysis of Wikipedia categories. However, comparing to Wikipedia with loose control of contributions, the LCSH ontology is believed more reliable, since it is developed by well trained category specialists, gone through continuous development for over a century (Chan 2005, Tao et al. 2011).

Feature extraction and selection techniques find informative, discriminative features to facilitate text categorisation. Related work includes Cai et al. (Cai et al. 2010), who exploited feature selection for unsupervised multi-clustering of images, with potential to extend the contribution to text categorisation. Forman and Kirshenbaum (Forman & Kirshenbaum 2008) introduced a fast method to extract text feature for categorisation. The work very much relies on text tokenisation and neglects the semantic concepts underlying from text; whereas semantic concepts are the focus in our work. Feature selection is also used by Yu et al. (Yu et al. 2010) to cluster documents via Dirichlet Process Mixture Model. In comparison, our work is different in sourcing global knowledge defined and specified in a world ontology, which helps to refine the result of feature extraction and selection. Malik and Kender (Malik & Kender 2008) proposed a pattern-based categorisation algorithm called "Democratic Classifier". Their work relies on the quality of training samples and cannot deal with the unsupervised problem. Having a common argument of that most of information on documents can be captured in phrases (or sub-strings), Hofmann et al. (Hofmann et al. 2009) studies the impact of document structure; Bekkerman and Matan (Bekkerman & Gavish 2011) investigate text categorisation independently. Their key phrases and sub-strings are in fact, sequential patterns. Furthermore, closed frequent patterns are used by Kiran et al. (Kiran et al. 2010) to help measure the distance between documents and clusters. Similarly, our work also extracts features by closed sequential frequent patterns, but again, with further investigation based on the mapping of local features to global ontological concepts.

3 Definitions

Text documents have some properties that make categorisation difficult. The structure and format are usually complex; the topics are heterogeneous and change with time. An efficient text categorisation approach needs to be able to handle these properties. Suggested by (Tao et al. 2011), an effective strategy is using world knowledge ontologies to guide the analysis of documents. Ontologies are formal descriptions and specifications of conceptualisation. By nature ontologies are a powerful instrument that can help to clarify and then solve complex, heterogeneous semantic problems.

The world knowledge ontology in this work is constructed based on the LCSH, as suggested by Tao et al. (Tao et al. 2011). The LCSH was developed for organising and retrieving information from a large volume of library collections. As pointed out by Chan (Chan 2005), the LCSH has many superiorities that can be taken to deal with the problems in text analysis:

- The LCSH system is an ideal world knowledge base covering an exhaustive range of topics. (Competent to deal with the complexity and heterogeneity issue);
- The LCSH represents the natural growth and distribution of human intellectual work. For over a hundred years, the knowledge contained in the LCSH has undergone continuous revision an enrichment. (Competent to deal with the topic change issue);
- The LCSH has the most comprehensive non-specialised controlled vocabulary in English. (Supplying an extensive category base for categorisation.)

The LCSH is also superior to other world knowledge ontologies. Tao et al. (Tao et al. 2011) compared the LCSH with the Library of Congress Classification, the Dewey Decimal Classification, and Yahoo! categorisation, and found that the LCSH is superior in topics, structure, and semantic relations. In many respects, the LCSH has become a de facto standard for subject cataloguing and indexing[1], and is used as a means for knowledge management systems (Chan 2005).

The concepts in the LCSH world ontology are called *subjects*. They are encoded from subject headings in the LCSH authorities. Subjects can be formalized:

Definition 1 *Let $\mathbb{S}$ be a set of subjects, an element $s \in \mathbb{S}$ is a 4-tuple $s := \langle label, neighbour, ancestor, descendant \rangle$, where*

- *label is the subject heading of s in the LCSH authorities, and $label(s) = \{t_1, t_2, \ldots, t_n\}$ where t is a term;*
- *neighbour is a function returning the subjects that have direct links to s in the LCSH authorities and $neighbour(s) \subset \mathbb{S}$;*
- *ancestor is a function returning the subjects that have higher level of abstraction than s and link to s directly or indirectly in the LCSH authorities and $ancestor(s) \subset \mathbb{S}$;*
- *descendant is a function returning the subjects that are more specific than s and link to s directly or indirectly in the LCSH authorities and $descendant(s) \subset \mathbb{S}$.* □

[1] Though the majority of libraries utilising the LCSH are located in United States, almost all libraries around the world have their systems convertible to the LCSH.

The semantic relations of subjects are encoded from the references defined in the LCSH authorities for subject headings, such as *Broader Term*, *Used for*, and *Related to*. The $ancestor(s)$ function in Definition 1 returns the *Broader Term* subjects of s (they are broader in semantics and thus, more general than s); the $descendant(s)$ returns the subjects that are *Used for* s and the subjects whose *Broader Term* is s (s is broader in semantics and thus, more general than these subjects); the $neighbour(s)$ returns the *Related to* subjects of s.

With Definition 1, the world knowledge ontology is defined:

Definition 2 *Let $\mathcal{O}$ be a world knowledge ontology. $\mathcal{O}$ consists of a set of subjects linked by their semantic relations, and can be formally defined as a 3-tuple $\mathcal{O} := \langle \mathbb{S}, \mathbb{R}, \mathbb{H}^{\mathbb{S}}_{\mathbb{R}} \rangle$.* □

where $\mathbb{R}$ is a set of relations specifying the relationships of subjects, as described previously; $\mathbb{H}^{\mathbb{S}}_{\mathbb{R}}$ is the taxonomical structure constructed by subjects in $\mathbb{S}$ linked by the relations in $\mathbb{R}$.

The study is to classify an unstructured document to a number of categories defined and specified in a world ontology. The study is limited to only English documents and focused on only the document content. To illustrate the research problem, we use a sample document, which is retrieved from the on-line catalogue of University of Melbourne Library[2]. The catalogue item forms a text document, which is composed by the catalogue title and content summary[3]:

> *Economic espionage and industrial spying. Dimensions of economic espionage and the criminalization of trade secret theft – Transition to an information society – increasing interconnections and interdependence – International dimensions of business and commerce – Competitiveness and legal collection versus espionage and economic crime – Tensions between security and openness – The new rule for keeping secrets - the Economic Espionage Act – Multinational conspiracy or natural evolution of market economy.*

By accessing to the catalogue item, a list of librarian manually-assigned subjects can be observed:

> *Business intelligence; Trade secrets; Computer crimes; Intellectual property; Commercial crimes.*

These title, summary, and subjects depict the ultimate goal we pursue in this work: given an unstructured document (e.g., the title and summary in the sample catalogue item), we are to categorise it into an indexed set of categories (subjects) specified in the world ontology (e.g., the listed subjects in the sample catalogue item). Ideally, the extracted subjects should be the same as these linguist manually-assigned subjects, because they are the result of human intellectual work. However, at this stage such a goal is unrealistic. Therefore, finding similar assignment of subjects with human work is the pursuing goal in this work.

With the research aim and scope, the research problem in this study can be formalized as the following function:

Definition 3 *Let $\Omega = \{i_1, i_2, \ldots, i_n\}$ be a finite and non-empty set of documents. Given an $i \in \Omega$, mappings need to be created for a set of subjects $\mathcal{S} \subseteq \mathbb{S}$:*

$$\eta : \Omega \to 2^{\mathcal{S}}, \quad \eta(i) = \{s \in \mathcal{S} | str(i,s) \geq min\_str\}$$

and also for its reverse mapping η^{-1}:

$$\eta^{-1} : \mathcal{S} \to 2^{\Omega}, \quad \eta^{-1}(s) = \{i \in \Omega | str(i,s) \geq min\_str\}$$

where $str(i,s)$ is the strength describing the validity rate of s to categorise i, and min_str is the threshold for the extent of qualified semantic space. □

4 Conceptual Framework

From Definition 3, given a document i, three elements, s, $str(i,s)$, and min_str, can be observed making impact to the categorisation results. Therefore, with respect to the elements, the research problem is decomposed into three tasks:

1. **Choosing a set of candidate subjects from the ontology to map to the document**;
2. **Investigating the relationships of subjects and the document and measure** $str(i,s)$;
3. **Evaluating** min_str **to find an adequate threshold.**

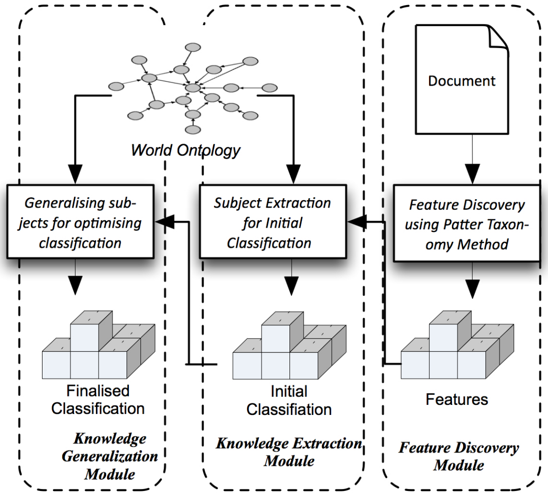
World Ontplogy
Document
Generalising subjects for optimising classification
Subject Extraction for Initial Classification
Feature Discovery using Patter Taxonomy Method
Finalised Classification
Initial Classification
Features
Knowledge Generalization Module
Knowledge Extraction Module
Feature Discovery Module

Figure 1: Conceptual Framework

Driven by the tasks, the conceptual framework of the proposed approach is designed consisting of three modules, as illustrated in Fig. 1.

Feature Extraction Module. Pattern Taxonomy Method is employed to discover features from the given document, based on the theory of closed frequent sequential patterns. The outcome would be a set of patterns with weights representing the features of document;

Knowledge Extraction Module. A term-subject matrix is established to extract appropriate subjects from the LCSH world ontology, based on the features extracted in *Feature Extraction Module*. The matrix has two attributes: the joint set of unique terms in the label of all subjects; the set of all subjects in the world ontology. As a result, given a set of patterns (features), a mapping set of subjects can be

[2] The catalogue item can be accessed by searching the title of "Economic Espionage and Industrial Spying" via http://cat.lib.unimelb.edu.au/search/t

[3] In this presented work we equally considered a document's title and summary as the content of the document. Treating them with different priorities will be investigated in our future work.

extracted, in which each subject is with a value indicating the strength of describing or discriminating the semantic content of document;

Knowledge Generalisation Module. The subjects extracted in *Knowledge Extraction Module* are investigated and generalised. The semantic relations with other subjects in the neighbourhood and the location in the ontological structure are studied. Finally, an refined indexed list of subjects are generalised as the categories assigned to the document.

5 Unsupervised Multi-label Categorisation

5.1 Feature Extraction from Document

Given a document $i = \{t_1, t_2, \ldots, t_m\} \in \Omega$, a set of features is first extracted, and defined by $\mathcal{F}(i) = \{\langle p, w(p)\rangle\}$, a set of valid patterns with weights. The patterns are extracted from the document using the closed frequent sequential pattern mining techniques. Pattern-based representation of features is considered superior to traditional lexicon-based representation (Li et al. 2010). Though lexicon features are easy to access by both human-users and computer systems, they usually contain many noisy features due to the sense ambiguity problem natually coming with terms, whereas pattern-based features has less noisy features because the representation considers the context of terms co-occurred in phrases. However, the pattern-based presentation has its own limit caused by the length of patterns. While longer patterns are more informative and discriminative, usually they are less frequent and as a result, have less evidence to represent the document content. In this paper, we define document features as weighted closed frequent sequential patterns. Algorithm 1 presents how the pattern features are discovered from a document.

```
input : $i = \{t_1, t_2, \ldots, t_n\}$, a threshold $\vartheta$.
output: The feature set $\mathcal{F}(i) = \{\langle p, w(p)\rangle\}$.
$\mathcal{F}(i) = \emptyset, p = \emptyset$;
for $(l = 1; l <= n; l++)$ do
    for $(j = l; j <= n; j++)$ do
        $p = p \cup \{t_j\}$;
        if $\exists\langle p', w(p')\rangle \in \mathcal{F}(i)$ where $p == p'$ then $w(p')++$;
        else $\mathcal{F}(i) = \mathcal{F}(i) \cup \{\langle p, 1\rangle\}$;
    end
end
foreach $\langle p, w(p)\rangle \in \mathcal{F}(i)$ do
    if $w(p) < \vartheta$ then $\mathcal{F}(i) = \mathcal{F}(i) - \{\langle p, w(p)\rangle\}$;
    else if $\exists\langle p', w(p')\rangle \in \mathcal{F}(i)$ where $(p \subset p') \wedge (w(p) \leq w(p'))$ then
    $\mathcal{F}(i) = \mathcal{F}(i) - \{\langle p, w(p)\rangle\}$
end
return $\mathcal{F}(i)$.
```

Algorithm 1: Feature Extraction from a Document

In practice, we employed the *pattern taxonomy model* (PTM) developed in by (Li et al. 2010) to discover the closed frequent sequential patterns from the document. The PTM model has been proven superior to other existing pattern mining approaches, such as *Probabilistic* and *Rocchio*, as reported in experimental evaluations. (The details of PTM model and its related evaluation can be found in (Li et al. 2010).) Thus, given a document i, $\mathcal{F}(i) = \{\langle p, w(p)\rangle\}$ is obtained, where $w(p)$ is the frequency of p in i. Table 1 displays the closed frequent sequential patterns extracted from the sample document presented in Section 3. Note the pattern discovery was performed with $min\_sup = 2$, and based on the text after word stemming and stopword removal.

Table 1: Closed Frequent Sequential Patterns Extracted From the Sample Document

Closed Sequential Pattern	Frequency
dimens	2
espionag	4
econom espionag	3
secret	2
econom	4

5.2 Mapping Local Features to Global Subjects

Let $\mathbb{T}$ be the term space of $\mathbb{S}$ in $\mathcal{O}$ and $\mathbb{T} = \bigcup_{s \in \mathbb{S}} label(s)$. A matrix coordinated by $\mathbb{T}$ and $\mathbb{S}$ is defined:

Definition 4 *Let $\langle \mathbb{S}, \mathbb{T}\rangle$ be the matrix coordinated by $\mathbb{T}$ and $\mathbb{S}$, where a mapping exists:*

$$\mu : \mathbb{T} \rightarrow 2^{\mathbb{S}}, \quad \mu(t) = \{s \in \mathbb{S} | t \in label(s)\} \subseteq \mathbb{S}$$

and its reverse mapping also exists:

$$\mu^{-1} : \mathbb{S} \rightarrow 2^{\mathbb{T}}, \ \ \mu^{-1}(s) = \{t \in \mathbb{T} | s \in \eta(t)\} \subseteq \mathbb{T}.$$ □

By $\mu : \mathbb{T} \rightarrow 2^{\mathbb{S}}$, a term $t \in \mathbb{T}$ maps to a set of subjects $\mathcal{S}_t \subseteq \mathbb{S}$. Thus, given the feature set $\mathcal{F}(i) = \{\langle p, w(p)\rangle\}$, a set of subjects can be extracted from $\mathbb{S}$:

$$\mathcal{S}_i = \bigcup_{t \in termset(\mathcal{F}(i))} \mu(t) \tag{1}$$

where $\mathcal{S}_i \subseteq \mathbb{S}$; $\mu(t) = \emptyset$ if $t \notin \mathbb{T}$.

By $\mu^{-1} : \mathbb{S} \rightarrow 2^{\mathbb{T}}$, a subject $s \in \mathbb{S}$ maps to a set of terms $\{t\} \subseteq \mathbb{T}$. Hence, with Eq. (1), a set of terms can be obtained from $\mu^{-1}(s)$ to expand i:

$$termset(i) = \bigcup_{s \in \mathcal{S}_i} \mu^{-1}(s) \tag{2}$$

Note that $termset(i) \neq i$. There exist some terms $\{t | t \in termset(i), t \notin i\}$, these terms are suggested by $\mathcal{S}_i$; there also exist some terms $\{t | t \notin termset(i), t \in i\}$ because those terms are not in the term space $\mathbb{T}$ and thus, map to an empty set of subjects.

Because $\mathcal{S}_i$ is extracted using $\mathcal{F}(i) = \{\langle p, w(p)\rangle\}$, considering the weights of feature patterns, we can evaluate $t \in termset(i)$:

$$w(t) = \sum_{p \in \{p | t \in termset(p), p \in \mathcal{F}(i)\}} w(p) \tag{3}$$

Considering the distribution of the terms spreading in other subjects, the normalized form of term evaluation is defined:

$$nw(t) = w(t) \times log(\frac{|\mathcal{S}_i|}{sf(t, \mathcal{S}_i)}) \tag{4}$$

where $sf(t, \mathcal{S}_i) = |\{s | t \in \mu^{-1}(s), s \in \mathcal{S}_i\}|$.

Subjects in $\mathcal{S}_i$ can finally be evaluated for their strengths of mapping i, using $nw(t)$ for all $t \in \mu^{-1}(s)$:

$$str(i, s) = \sum_{t \in \mu^{-1}(s)} nw(t) \tag{5}$$

By using the normalised form of terms, the subjects are competent for not only describing i but also discriminating i from other documents in Ω.

Table 2: Subjects Representing the Sample Document

Subject	**Strength**
Espionage	16.83
Espionage, economic	13.01
Space surveillance	13.01
Dimensions	9.24
Espionage, industry	9.24
Business espionage	8.98
Espionage literature	8.98
Espionage story	8.98
… …	…

In order to prune away noisy subjects, the threshold, *min_str*, is applied to subject extraction. The subjects with $str(i, s) \geq min\_str$ are kept; whereas those with $str(i, s) < min\_str$ are dropped. During the experiments, different values were evaluated for *min_str*. The results revealed that setting *min_str* as the top fifth $str(i, s)$, a variable rather than a static value, gave the system the best performance. Therefore, the *min_str* was set up dynamically as the top fifth value of $str(i, s)$ for subject selection. Table 2 shows the valid subjects extracted from the LCSH ontology for the sample document in Section 3, using the closed frequent sequential patterns shown in Table 1. Note that only the top subjects are displayed, because there are a total of 80 subjects survived the pruning process.

5.3 Generalising Subjects for Optimal Categorisation

The subject set extracted from the ontology (as described in Section 5.2) suffers from a problem - the subject set is easily oversized. For example, there are 80 subjects selected for the sample document in Section 3. As a result, the system's complexity becomes high, and performance becomes difficult to handle, when the category (subject) base has a large volume. The extracted subject set needs to be generalised and optimised for better categorisation.

Some subjects extracted from the ontology are observed overlapping in their semantic space. For example, for the subjects displayed in Table 2, by common sense we know that 'Espionage' dominates 'Espionage, economic', 'Espionage, industrial', and 'Business espionage'; that 'Espionage literature' dominates 'Espionage story'. This is caused by the same feature terms occurred in their labels. This observation raises an idea of generalising the initial subject set by pruning away overlapping subjects.

This generalisation is accomplished via investigating the relations existing between subjects. From Definitions 1 and 2 we know subjects in $\mathcal{O}$ have semantic relations linking each other. A subject s may have $ancestor(s)$, a set of subjects linking to s with higher level of abstraction than s; and $descendant(s)$, a set of linking subjects more specific than s. Such a taxonomical structure is constructed based on the semantic extent and focus of subjects. Let s_1 and s_2 be two subjects and $s_1 \in ancestor(s_2)$ ($s_2 \in descendant(s_1)$). s_1 refers to a larger semantic extent than s_2 and thus, is more general than s_2. On the other hand, s_2 is more specific than s_1 and thus, more focuses on its referring-to concept. Such relationships can be revealed from an example of s_1 as 'Auto-mobile' and s_2 'Sedan'. 'Auto-mobile' contains 'Car', 'Truck', *etc*, and 'Car' contains 'Sedan', 'Hatchback', etc. 'Auto-mobile' covers broader extent than 'Sedan'; vise versa, 'Sedan' is more focused than 'Auto-mobile'. Therefore, in the extracted subject set, if one subject is an descendant of another, the descendant can be removed because its referring-to semantic extent has been covered by its ancestor. By removing the descendant subjects, we have no information loss but just focus, for example, replacing 'Sedan' by 'Car' if they both in the extracted subject set.

Following the same rule, if some subjects are under the same umbrella of an ancestor, they can be replaced by this common ancestor without information loss, though the common ancestor is not in the extracted subject set. Thus, if 'Sedan' and 'Hatchback' are in the set, they may be replaced by their common ancestor 'Car'. We may also use 'Auto-mobile' to replace 'Sedan' and 'Truck', though they are not at the same taxonomic level in the ontology. However, the common ancestor chosen to replace its descendant subjects cannot be too far away from the replaced descendants in the taxonomic structure. One extreme example is that we cannot use 'Thing' to replace any subjects, as 'Thing' is the root and dominates all subjects in the ontology. An ancestor subject being too far from its descendants becomes meaningless to them, because the focus is severely lost. Therefore, we use only the lowest common ancestor (shortened by $\mathcal{LCA}$) to replace its descendant subjects. The lowest common ancestor is a proven effective technique to help capture semantic meaning in text. Nguyen and Cao (Nguyen & Cao 2010) exploited a variant of $\mathcal{LCA}$, namely Relevant Lowest Common Ancestor, to capture accurate relevant fragments in XML search. In this work, the $\mathcal{LCA}$ refers to the common ancestor of a set of subjects, with the shortest distance to these subjects in the taxonomic structure of ontology. The $\mathcal{LCA}$ dominates these descendant subjects and covers their semantic extent with only limited loss of focus. Replacing these descendant subjects by their $\mathcal{LCA}$ generalises the semantic content.

```
    input : $\mathcal{S}_i = \{s_1, s_2, \ldots, s_j\}$ (subject set extracted $i$),
            $\mathcal{O}$;
    output: $\mathcal{S}'_i = \{s_1, s_2, \ldots, s_k\}$ (subject set generalized
            to map $i$).
1   $\mathcal{S}'_i = \emptyset, \mathcal{S}_{temp} = \emptyset, \mathcal{S}_{redundant} = \emptyset$;
2   foreach $s \in \mathcal{S}_i$ do
3       Extract $S(s)$ from $\mathcal{O}$ where
        $S(s) = \{s' | s' \in ancestor(s), \delta(s \mapsto s') \leq 3\}$;
        foreach $s_n \in \mathcal{S}_i$ where $s_n \neq s$ do
4           Extract $S(s_n)$ from $\mathcal{O}$ like Step 3;
5           if $S(s) \cap S(s_n) \neq \emptyset$ then
            $\{\widehat{s} = \mathcal{LCA}(S(s) \cup S(s_n)), str(i, \widehat{s}) =$
            $str(i, s) + str(i, s_n); \mathcal{S}_{temp} = \mathcal{S}_{temp} \cup \{\widehat{s}\};$
            $\mathcal{S}_{redundant} = \mathcal{S}_{redundant} \cup \{s, s_n\}\}$
6       end
7       if $\mathcal{S}_{temp} \neq \emptyset$ then $\{\mathcal{S}'_i = \mathcal{S}'_i \cup \mathcal{S}_{temp}$;
        $\mathcal{S}_i = \mathcal{S}_i - \mathcal{S}_{redundant}$; $\mathcal{S}_{temp} = \emptyset$; $\mathcal{S}_{redundant} = \emptyset\}$
        else $\mathcal{S}'_i = \mathcal{S}'_i \cup \{s\}$
8   end
9   return $\mathcal{S}'_i$.
```

Algorithm 2: Generalizing Subjects

Algorithm 2 explains how to generalise a set of subjects to optimise the categorisation of a document. The function $\delta(s_1 \mapsto s_2)$ returns a positive real number indicating the distance between two subjects. The distance is measured by counting the number of edges travelled through from s_1 to s_2 in the taxonomic structure of $\mathcal{O}$. The function $\mathcal{LCA}(S(s_1) \cup S(s_2))$ returns $\widehat{s}$, the $\mathcal{LCA}$ of s_1 and s_2 in a joint subject set, $S(s_1) \cup S(s_2)$. Note that $\delta(s_1 \mapsto s_2) \leq 3$, which means only the $\mathcal{LCA}$s within the distance of three edges are considered. Subjects further than that in distance are too general; whereas using a highly-general subject for generalisation would severely loose the focus

Table 3: Generalized Subjects for the Sample Document

Subject	Strength
Espionage	269.78
Business Intelligence	203.83
Space surveillance	17.96
Spy story	16.27
Dimensions	9.24

of original subjects. As a result, the original, specific information in the document is jeopardised. (In our experiments, $\delta(s_1 \mapsto s_2) \leq 3$ and ≤ 5 were tested under the same environment in order to find a valid distance for tracking the competent $\mathcal{LCA}$. The system achieved a better performance when $\delta(s_1 \mapsto s_2) \leq 3$. The details of this sensitivity test can be found in Section 7.)

Table 3 presents the subjects generalised from those displayed in Table 2. Again, the *min_str* is set as the top fifth $str(i, s)$ value. Similar subjects, for example, 'Espionage', 'Espionage, economic', 'Espionage, industrial', and 'Business espionage', have been merged and replaced by their $\mathcal{LCA}$. 'Espionage' and 'Business Intelligence'; 'Espionage literature' and 'Espionage story' have been replaced by 'Spy story'. Consequently, the 80 subjects initially extracted from the ontology (as described in Section 5.2 previously) are generalised to a much shorter list containing only five subjects. This set of subjects is more applicable to the systems, and more competent for the categorisation of the sample document presented in Section 3.

6 Evaluation

6.1 Experiment Design

Ideally, to categorise a document, the subjects automatically generated by the proposed approach should be exactly the same as those specified by specialist librarians. Though such a goal is unrealistic, the ideal scenario inspirited the design of our evaluation experiments - the proposed method was evaluated, based on the ground truth of manual assignment of subjects and compared with typical categorisation approaches. The experiments were performed using a large corpus obtained from the catalogue of a library employing the LCSH authorities. A sample catalogue item has been presented in Section 3. The text of each item in the catalogue was parsed to form a document. The title and content of catalogue items were used to form the content. The subject headings assigned to the catalogue items were manually specified from the LCSH authorities by specialist librarians who were trained to specify subjects for a document without bias (Chan 2005). They provided the ideal ground truth in the experiments to measure the effectiveness of the proposed approach, against the automatically generated subjects. By using the catalogue items in a library as the corpus, we could easily obtain a large testing set as well as a perfect ground truth for evaluation. The objective evaluation methodology also assured the solidity and reliability of the experimental evaluation for our proposed method.

The testing set was retrieved from the on-line catalogue of the University of Melbourne Library[4] by using as queries the title of topics (R101-R150) created by the TREC-11 Filtering Track. These topics were manually designed by linguists for evaluation of information retrieval methods. The retrieved catalogue items were parsed to keep only title, content, and subjects. The testing set was then generated by pooling the catalogue items retrieved by total 50 topics. Text pre-processing techniques, such as stopword removal and word stemming (Porter stemming algorithm), were applied to the preparation of testing set for experiments. Table 4 shows the statistics of the testing set. In the experiments, we used only documents with at least 30 terms after stopword removal. Documents shorter than that could hardly provide substantial frequent patterns, as revealed in the experiments.

Description	Stat.
Documents crawled	227,219
Documents used in experiments	31,902
Shortest document in experiments	30
Longest document in experiments	952
Average length of documents in exps	85

Table 4: Statistics of the Testing Set

6.2 Baseline Models

Given that the LCSH ontology contains 394,070 subjects in our implementation, the category base has a fairly large volume. Hence, we chose two typical multi-class categorisation approaches, *Rocchio* and kNN, as the baseline models in the experiments.

Rocchio is an efficient classification method using centroid to define the class boundaries. The centroid of a class c is computed as the vector average:

$$\vec{\mu}(c) = \frac{1}{|D_c|} \sum_{i \in D_c} \vec{v}(i) \qquad (6)$$

In the experiments, a class c referred to a subject s. The training set D_c contained only a single document $i = label(s)$. The distance between a document and a subject was measured by cosine similarity. The document was then categorised into the subject classes with the highest cosine value. (Considering that the category base volume is a huge number, using only the top cosine value has already produced a considerably large set of subjects.)

Unlike *Rocchio*, k Nearest Neighbour (kNN) determines the decision boundary locally and categories documents into the major categories of its k closest neighbours. When incoming a document d, from the testing set we extracted the closest neighbours $NN(d)$ that had the highest cosine similarity value with d. Because the testing documents were usually short, a large number of documents were found having the same cosine values. Thus, we set $k = 1$ to limit the number of considerable neighbours, as well as to ensure the highest possible accuracy. The distance of a subject s and a document d is then evaluated by aggregating the cosine value of each $d' \in NN(d)$ to s. Once again, d was classified into the subjects with the highest cosine similarity value.

6.3 Performance Measuring Methods

The performance of experimental models was measured by precision and recall, the modern evaluation methods in information retrieval (IR). Precision was to measure the ability of a method categorising a document into highly focusing categories, and recall the ability of categorising a document into categories without missing any potential ones.

[4] http://www.library.unimelb.edu.au/

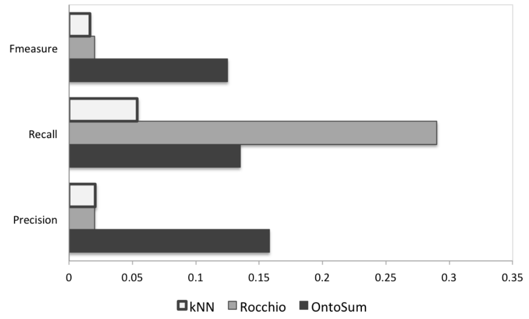

Figure 2: Effectiveness Performance on Average

As discussed previously, considering the category base volume, seeking exact-mapping of subjects was unrealistic. Thus, in respect with the text corpus and the ground truth featured by the LCSH, the system performance was evaluated by:

$$precision = \frac{|FT(S_{tgt}) \cap FT(S_{grt})|}{|FT(S_{tgt})|} \tag{7}$$

$$recall = \frac{|FT(S_{tgt}) \cap FT(S_{grt})|}{|FT(S_{grt})|} \tag{8}$$

where $FT(S) = \bigcup_{s \in S} \mu^{-1}(s)$ (see Definition 4); tgt referred to the target experimental model; grt referred to ground truth subjects.

F_1 Measure as another common method used in IR was also employed in our experiments:

$$F_1 = \frac{2 \times precision \times recall}{precision + recall} \tag{9}$$

Precision and recall were evenly considered in F_1 Measure. In this evaluation, we used *micro*-F_1 Measure, which evaluated each document's mapping result first and then averaged the evaluation results for the final F_1 Measure value. Greater F_1 values indicate better performance.

7 Results and Discussions

7.1 Experimental Results

By calling the proposed approach as the OntoMap model, the experiments were to compare the effectiveness performance of the OntoMap model to the baselines, Rocchio and *k*NN models. Their effectiveness performances are depicted in Fig. 3, 4, and 5 for the number of documents with valid effectiveness (> 0), where the value axis indicates the effectiveness rate between 0 and 1; the category axis indicates the number of documents whose mappings meet the indicating effectiveness rate. As shown in the figure and discussed in Section 6.3, the effectiveness rates are measured by precision, recall, and F_1 Measure, where $P(X)$ refers to the precision results of experimental model X, $R(X)$ the recall results, and $F(X)$ the F_1 Measure results. Their overall average performances are shown in Fig. 2.

F_1 Measure equally considers both precision and recall in performance measuring. Thus the F_1 Measure results can be deemed as an overall effectiveness performance. The average F_1 Measure result illustrated in Fig. 2 reveals that the OntoMap model has achieved a much better overall performance than the two baseline models. The performance is also confirmed by the detailed results depicted in Fig. 3 - the $F(OntoMap)$ line is located at much higher bound level compared to the $F(Rocchio)$ and $F(kNN)$ lines.

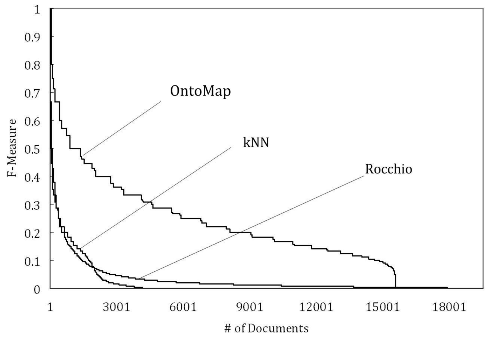

Figure 3: Performance Measured by F_1 Measure

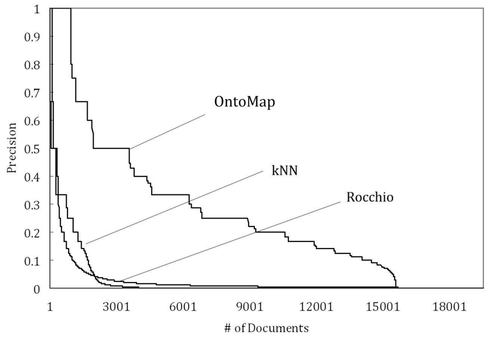

Figure 4: Performance Measured by Precision

Precision measures how effective the global and local knowledge mappings are. In terms of that, the OntoMap model once again has outperformed the baseline models. The average precision results shown in Fig. 2 demonstrates this achievement. The detailed precision results depicted in Fig. 4 illustrate the same message that the $P(OntoMap)$ curve is depicted much higher than those of the baseline models.

Recall measures the rate of global and local knowledge mappings covering all dealing-with subjects. The recall performance shows a slightly different result compared to those from F_1 Measure and precision performance. The *Rocchio* model achieved the best recall performance, compared to that of the OntoMap model and the *k*NN model. This message is also illustrated in Fig. 5 as $R(OntoMap)$ lies in the middle of $R(Rocchio)$ and $R(kNN)$.

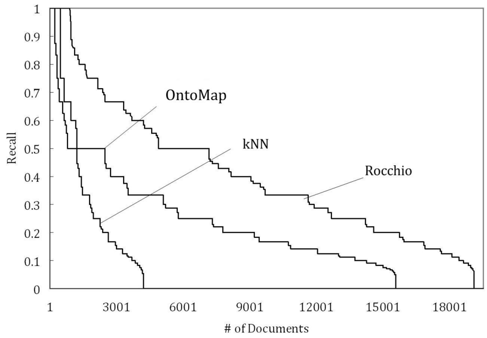

Figure 5: Performance Measured by Recall

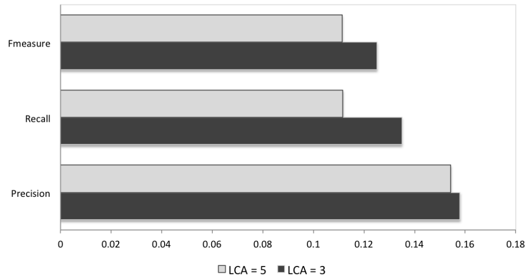

Figure 6: Effectiveness Performance on Average

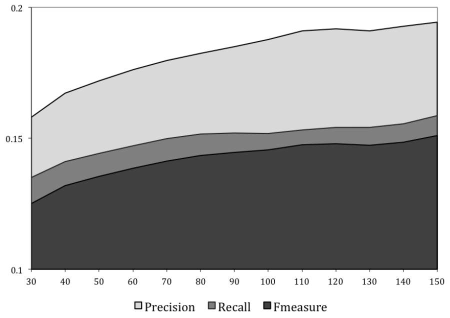

Figure 7: OntoMap Performance on Documents with Different Lengths

7.2 Discussions

There was a gap between the recall performance of the OntoMap and the Rocchio models. After the investigation conducted into the insight of recall results, we found that the categorisation made by the Rocchio model was usually a large set of subjects (935 on average), whereas the OntoMap model had results with a much reasonable size of subject sets (16 on average) and the kNN's had an average size of 106. Due to the nature of recall measurement, more feature term would be covered if the subject size became larger. As a result, the Rocchio mappings with the largest size of subjects achieved the best recall performance. The category sets generated by the kNN model had larger size than those of the OntoMap. However, when taking neighbours into consideration, a large number of nosey subjects was also brought into the neighbourhood consideration - the average number of neighbours was 336. This was caused by the extremely high dimension in the category base and short length of documents (average length=85) in experiments. As a result, the multi-label categorisation became ineffective because only the documents with top cosine values were chosen to expand and only the subjects with the top similarity values were chosen for categories.

Different values were tested in sensitivity study for choosing a right number of levels to find the lowest common ancestor when generalising subjects for final mappings. (The related discussion can be referred to Section 5.3.) Fig. 6 displays the testing results for finding such a right level. In the same experimental environment, if tracing three levels to find a $\mathcal{LCA}$ the OntoMap model's overall performance including F_1 Measure, precision, and recall was better than that by five levels. In addition, tracing only three levels costs less in complexity compared with five levels. Therefore, we chose three levels to restrict the extent of finding $\mathcal{LCA}$s.

We also found that the performance of the OntoMap model was improved when longer documents were used in the experiments. Figure 7 depicted the improvement made by the OntoMap model by restricting documents with a minimum length set from 30 till 150. When the minimum length was set to 30, about 32,000 documents were tested, the performance was the same as that discussed in Section 7.1; when set to 150, the documents dropped to 2,650, and the performance was much better. Figure 7 reveals that the effectiveness of categorisation increases when the length of considering documents increases. Such an improvement is contributed by the closed frequent sequential patterns discovered for document feature representation (see Section 5.1 for details). When the OntoMap has the best performance with restricting to only documents longer than 150 terms, the average number of closed frequent sequential patterns was 27; when restricting to documents with length>= 90, the average number of patterns dropped to 17; when with all the documents (length>= 30), the average number of discovered patterns dropped to 11. These facts reveal that more meaningful patterns are discovered from long documents with more semantic contents.

8 Conclusions and Future Work

Text categorisation has been widely exploited to assist tasks in information retrieval, information organising, text categorisation, and knowledge engineering. Traditionally, text categorisation relies on the quality of training samples with category labels, the informative, discriminative features extracted from documents, and the topic coverage of category base. Sometimes qualified training samples may be absent, the problem then becomes unsupervised. When one single category assigned to the document cannot not fully describe the content, the single label categorisation then becomes a multi-label problem. In this paper we have introduced an unsupervised multi-label text categorisation approach exploiting a world ontology. The ontology is encoded from the Library of Congress Subject Headings and contains a large volume of subjects with extensive topic coverage. The proposed approach uses the subjects as the category base, and maps the local features extracted from the document to a set of global categories. The approach then studies the semantic relations and ontological structure to optimise the categorisation result to the document. The approach has been experimentally evaluated by comparing with typical methods including *Rocchio* and kNN , using a large real-world corpus, based on the ground truth encoded by human experts. The experimental results demonstrated that the OntoMap model outformed baseline models of Rocchio and kNN, in terms of overall effectiveness and precision performance.

The work presented in this paper holds some limitations that need to breakthrough in our future work. Though the proposed approach is able to deal with relatively short text documents (with minimum number of 30 terms, as demonstrated in experiments), it is incapable of handling extremely short documents because of using frequent sequential patterns. However, the Web, especially Web 2.0, has a large portion of extremely short documents (e.g., tweets are limited to 140 characters). To handle these texts the proposed method still needs to be improved. The current work also largely relies on the world knowledge ontology that can only be updated with the LCSH. How to update the ontology efficiently also remains as an open question for us to pursue in the future.

References

Bekkerman, R. & Gavish, M. (2011), High-precision phrase-based document classification on a modern scale, *in* 'Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining', pp. 231–239.

Cai, D., Zhang, C. & He, X. (2010), Unsupervised feature selection for multi-cluster data, *in* 'Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining', pp. 333–342.

Camous, F., Blott, S. & Smeaton, A. F. (2007), Ontology-based medline document classification, *in* 'Proceedings of the 1st international conference on Bioinformatics research and development', pp. 439–452.

Chan, L. M. (2005), *Library of Congress Subject Headings: Principle and Application*, Libraries Unlimited.

Cheng, W., Kasneci, G., Graepel, T., Stern, D. & Herbrich, R. (2011), Automated feature generation from structured knowledge, *in* 'Proceedings of the 20th ACM international conference on Information and knowledge management', pp. 1395–1404.

Forman, G. & Kirshenbaum, E. (2008), Extremely fast text feature extraction for classification and indexing, *in* 'Proceedings of the 17th ACM conference on Information and knowledge management', pp. 1221–1230.

Gabrilovich, E. & Markovitch, S. (2005), Feature generation for text categorization using world knowledge, *in* 'Proceedings of The Nineteenth International Joint Conference for Artificial Intelligence', pp. 1048–1053.

Hernandez, N., Mothe, J., Chrisment, C. & Egret, D. (2007), 'Modeling context through domain ontologies', *Information Retrieval* **10**, 143–172.

Hofmann, K., Tsagkias, M., Meij, E. & de Rijke, M. (2009), The impact of document structure on keyphrase extraction, *in* 'Proceedings of the 18th ACM conference on Information and knowledge management', pp. 1725–1728.

Hu, X., Zhang, X., Lu, C., Park, E. K. & Zhou, X. (2009), Exploiting wikipedia as external knowledge for document clustering, *in* Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining', pp. 389–396.

Kiran, G. V. R., Shankar, R. & Pudi, V. (2010), Frequent itemset based hierarchical document clustering using wikipedia as external knowledge, *in* 'Proceedings of the 14th international conference on Knowledge-based and intelligent information and engineering systems: Part II', pp. 11–20.

Li, Y., Algarni, A. & Zhong, N. (2010), Mining positive and negative patterns for relevance feature discovery, *in* 'Proceedings of 16th ACM SIGKDD Conference on Knowledge Discovery and Data Mining', pp. 753–762.

Malik, H. H. & Kender, J. R. (2008), Classifying high-dimensional text and web data using very short patterns, *in* 'Proceedings of the 8th IEEE International Conference on Data Mining', pp. 923–928.

Nguyen, K. & Cao, J. (2010), Exploit keyword query semantics and structure of data for effective xml keyword search, *in* 'Proceedings of the 21st Australasian Conference on Database Technologies - Volume 104', pp. 133–140.

Sun, X., Wang, H., Li, J. & Pei, J. (2011), 'Publishing anonymous survey rating data', *Data Mining and Knowledge Discovery* **23(3)**, 379–406.

Sun, X, Wang, H. Li, J. & Zhang, Y. (2012), 'Satisfying privacy requirements before data anonymization', *The Computer Journal* **55(4)**, 422–437.

Tao, X., Li, Y. & Zhong, N. (2011), 'A personalized ontology model for web information gathering', *IEEE Transactions on Knowledge and Data Engineering,* **23**(4), 496–511.

Tao, X., Li, Y., Zhong, N. & Nayak, R. (2008), 'A knowledge retrieval model using ontology mining and user profiling', *Integrated Computer-Aided Engineering* **15**(4), 313–329.

Tjondronegoro, D., Tao, X., Sasongko, J. & Lau, C. H. (2011), Multi-model summarization of key events and top players in sports tournament videos, *in* 'Proceedings of the IEEE Workshop on Applications of Computer Vision', pp. 471–478.

Wang, P. & Domeniconi, C. (2008), Building semantic kernels for text classification using wikipedia, *in* Proceeding of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining', pp. 713–721.

Yan, Y., Okazaki, N., Matsuo, Y., Yang, Z. & Ishizuka, M. (2009), Unsupervised relation extraction by mining wikipedia texts using information from the web, *in* 'Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Volume 2, pp. 1021–1029.

Yang, B., Sun, J.-T., Wang, T. & Chen, Z. (2009), Effective multi-label active learning for text classification, *in* 'Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining', pp. 917–926.

Yang, T., Jin, R., Jain, A. K., Zhou, Y. & Tong, W. (2010), Unsupervised transfer classification: application to text categorization, *in* 'Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining', pp. 1159–1168.

Yu, G., Huang, R. & Wang, Z. (2010), Document clustering via dirichlet process mixture model with feature selection, *in* 'Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining', pp. 763–772.

Zhong, N., Li, Y. & Wu, S.-T. (2012), 'Effective Pattern Discovery for Text Mining', *IEEE Transactions on Knowledge and Data Engineering,* **24**(1), 30–44.