Unsupervised Multi-Label Text Classification Using a World Knowledge Ontology

Xiaohui Tao[1], Yuefeng Li[2], Raymond Y. K. Lau[3], and Hua Wang[1]

[1]Centre for Systems Biology, University of Southern Queensland, Australia
[2]Science and Engineering Faculty, Queensland University of Technology, Australia
[3]Department of Information Systems, City University of Hong Kong, Hong Kong
[1]{xtao, hua.wang}@usq.edu.au, [2]y2.li@qut.edu.au, [3]raylau@cityu.edu.hk

Abstract. The development of text classification techniques has been largely promoted in the past decade due to the increasing availability and widespread use of digital documents. Usually, the performance of text classification relies on the quality of categories and the accuracy of classifiers learned from samples. When training samples are unavailable or categories are unqualified, text classification performance would be degraded. In this paper, we propose an unsupervised multi-label text classification method to classify documents using a large set of categories stored in a world ontology. The approach has been promisingly evaluated by compared with typical text classification methods, using a real-world document collection and based on the ground truth encoded by human experts.

1 Introduction

The increasing availability of documents in the past decades has greatly promoted the development of information retrieval and organising systems, such as search engines and digital libraries. The widespread use of digital documents has also increased these systems' accessibility to textual information. A fundamental theory supporting these information retrieval and organising systems is that information can be associated with semantically meaningful categories. Such a theory supports also ontology learning, text categorisation, information filtering, text mining, and text analysis, etc. Text classification aims at associating textual documents with semantically meaningful categorises, and has been studied in the past decades, along with the development of information retrieval and organising systems [11].

Text classification is the process of classifying an incoming stream of documents into predefined categories. Text classification usually employs a supervised learning strategy with the classifiers learned from pre-classified sample documents. The classifiers are then used to classify incoming documents. In terms of supervised text classification, the performance is determined by the accuracy of pre-classified training samples and the quality of the categorisation. The accuracy of classifiers determines their capability of differentiating the incoming stream of documents; the descriptive and discriminative capacity of categorisation reduces noise in classification, which is caused by sense ambiguities, sparsity, and high dimensionality of the documents [7]. Text classification performance is also affected by the topic coverage of categories. An inadequate category may be assigned to a document if an in-comprehensive set of categories is employed, because non-adequate categories can be found. The performance of text classification relies upon the descriptive and discriminative capacity of categories and the accuracy of classifiers learned from training sets.

However, there exist situations that a qualified training document set may not be available (e.g., the "cold start" problem in recommender systems); a set of categories with in-comprehensive topic coverage may be used for classification; sometimes although a set of categories with comprehensive topic coverage is available, the large number of classes would easily introduce noise in classification results [5]. Traditionally, text classification models are designed to handle only single-label problems. However, in some circumstances (e.g., categorizing documents in library catalogue into multiple subjects), multi-label text classification is required and automatic classification is necessary, especially when classifying a very large volume of documents [15]. To deal with these problems, in this paper we propose an automatic unsupervised text classification approach to classify documents into multiple classes, without the requirement of pre-classified sample documents for training classifiers. The approach consists of three modules; pattern mining for document feature extraction; feature-subject mapping for initial classification; knowledge generalisation for optimal classification. The method incorporates comprehensive world knowledge stored in a large ontology and classifies documents into the classes in the ontology without any pre-classified training samples available. The world ontology is built from Library of Congress Subject Headings (LCSH), which represents the natural growth and distribution of human intellectual work [4]. The subject classes and semantic relationships in the ontology are investigated and exploited to improve the classification results. The proposed method was experimentally evaluated using a large library catalogue, by compared with typical text classification approaches. The presented work makes three-fold contributions:

- An unsupervised text classification method that classifies documents into multiple classes;
- A knowledge generalisation method to optimise text classification by analysing the semantic relations of categories;
- An exploration of using the LCSH as a world knowledge to facilitate text classification.

The paper is organised as follows. Section 2 discusses the related work; Section 3 introduces the research problem and the the conceptual model of proposed supervised text classification method; Section 4 presents the technical detail of the proposed method. The experiment design is described in Section 5, whereas the results are discussed in Section 6. Finally, Section 7 makes conclusions.

2 Related Work

Unsupervised text classification aims to classify documents into the classes with absence of any labelled training documents. In many occasions the target classes may not have any labelled training documents available. One particular example is the "cold start" problem in recommender systems and social tagging. Unsupervised classification can automatically learn an annotation model to make recommendations or label the tags when the products or tags are rare and do not have any useful information associated. Unsupervised classification has been studied by many groups and many successful models have been proposed. Without associated training samples, Yang et al. [16] built a classification model for a target class by analysing the correlating auxiliary classes. Though as similar as theirs in investigating correlating classes, our work is different by exploiting a hierarchical world knowledge ontology for classification, instead of only auxiliary classes. Also exploiting a world knowledge base, Yan et al. [14] examined unsupervised relation extraction from Wikipedia articles and integrated linguistic analysis with web frequency information to improve unsupervised classification performance. However, our work has different aims from theirs; ours aims to exploit a world knowledge ontology to help unsupervised classification, whereas Yan et al. [14] aims to extract semantic relations for Wikipedia concepts by using unsupervised classification techniques. Cai et al. [2] and Houle and Grira [6] proposed unsupervised approaches to evaluate and improve the quality of selecting features. Given a set of data, their work is to find a subset containing the most informative, discriminative features. Though the work presented in this paper also relies on features selected from documents, the features are further investigated with their referring-to ontological concepts to improve the performance of classification.

Text classification models are originally designed to handle only single-label problems, where each document is classified into only one class. However, in many circumstances single-label text classification cannot satisfy the demand, for example, in social network multiple labels may need to be suggested for a tag [8]. Comparing with the work done by Katakis et al. [8], our work relies on the semantic content of documents, rather than the meta-data of documents used in [8]. As similar as the work conducted by Yang et al. [15], our work also targets on multi-label text classification. However, Yang et al. [15]' work is different in adopting active learning algorithms for multi-label classification, whereas ours exploits concepts and their structure in world knowledge ontologies.

Ontologies have been studied and exploited by many works to facilitate text classification. Gabrilovich and Markovitch [5] enhanced text classification by generating features using domain-specific and common-sense knowledge in large ontologies with hundreds of thousands of concepts. Comparing with their work, our work moves beyond feature discovery and investigates the hierarchical ontology structure for knowledge generalisation to improve text classification. Camous et al. [3] also introduced a domain-independent method that uses the Medical Subject Headings (MeSH) ontology. The method observes the inter-concept relationships and represents documents by MeSH subjects. Similarly, Camous' work considers the semantic relations existing in the ontological concepts. However, their work focuses on only the medical domain, whereas our approach works on general areas because exploiting the LCSH, a superior world knowledge ontology. Another world ontology commonly used in text classification is Wikipedia. Wang and Domeniconi [13] and Hu et al. [7] derived background knowledge from Wikipedia to represent documents and attempted to deal with the sparsity and high dimensionality problems in text classification. Instead of Wikipedia with free-contributed entries, our work uses the superior LCSH ontology, which has been under continuous development for a hundred years by knowledge engineers.

Many works utilise pattern mining techniques to help build classification models, which is similar as the strategy employed in our work. Malik and Kender [10] proposed the "Democratic Classifier", which is a pattern-based classification algorithm using short patterns. Different from our work, their democratic classifier relies on the quality of training samples and cannot deal with the "no training set available" problem. Bekkerman and Matan [1] argued that most of information on documents can be captured in phrases and proposed a text classification method that employs lazy learning from labelled phrases. The phrases in their work are in fact a special form of sequential patterns that are used in our work for feature extraction of documents.

3 Unsupervised Multi-label Text Classification

Let $\mathcal{D} = \{d_i \in \mathbb{D}, i = 1, \ldots, m\}$ be a set of text documents; $\mathcal{S} = \{s_1, \ldots, s_K\}$ be a large set of classes, where K is the number of classes. If there is available a training set $\mathcal{D}_t = \{d_j \in \mathbb{D}, j = m+1, \ldots, n\}$ with $y_j^k = \{0, 1\}, k = 1, \ldots, K$ provided for describing the likelihood of d_j belonging to class s_k, it is easy to learn a binary prediction function $p(y^k|d)$ and use it to classify $d_i \in \mathcal{D}$. However, our objective is to learn a prediction function $p(y^k|d)$ to classify d_i into $\{s_k\} \subset \mathcal{S}$ without $\mathcal{D}_t$ available. We refer to this problem as *unsupervised multi-label text classification.*

The proposed classification method consists of three steps: feature extraction, initial classification, and optimising classification, using a world ontology.

3.1 World Ontology

The world knowledge ontology is constructed from the Library of Congress Subject Headings (LCSH), which is a knowledge system developed for organising information in large library collections. It has been under continuous development for over a hundred years to describe and classify human knowledge. Because of the endeavours dedicated by the knowledge engineers from generation to generation, the LCSH has become a de facto standard for concept cataloguing and indexing, superior to other knowledge bases. Tao et al. [12] once compared the LCSH with the Library of Congress Classification, the Dewey Decimal Classification, and Yahoo! categorisation, and reported that the LCSH has broader topic coverage, more meaningful structure, and more accurate semantic relations. The LCSH has been widely used as a means for many knowledge engineering and management works [4]. In this work, the class set $\mathcal{S} = \{s_1, \ldots, s_K\}$ is encoded from the LCSH subject headings.

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

- *label is a set of sequential terms describing s; $lable(s) = \{t_1, t_2, \ldots, t_n\}$;*
- *neighbour refers to the set of subjects in the LCSH that directly link to s, $neighbour(s) \subset \mathcal{S}$;*
- *ancestor refers to the set of subjects directly and indirectly link to s and locating at more abstractive level than s in the LCSH, $ancestor(s) \subset \mathcal{S}$;*
- *descendant refers to the set of subjects directly and indirectly link to s and locating at more specific level than s in the LCSH, $descendant(s) \subset \mathcal{S}$.* □

The semantic relationships of subjects are encoded from the references defined in the LCSH for subject headings, including *Broader Term*, *Used for*, and *Related to.* The $ancestor(s)$ in Definition 1 returns the *Broader Term* subjects of s; the $descendant(s)$ is the reversed function of $ancestor(s)$, with additional subjects *Used for s*; the $neighbour(s)$ returns the subjects *Related to s*.

With Definition 1, the world knowledge ontology is defined:

Definition 2. (ONTOLOGY) *Let $\mathcal{O}$ be a world ontology. $\mathcal{O}$ contains a set of subjects linked by their semantic relations in a hierarchical structure. $\mathcal{O}$ is a 3-tuple $\mathcal{O} := \langle \mathcal{S}, \mathcal{R}, \mathcal{H}^{\mathcal{S}}_{\mathcal{R}} \rangle$, where*

- *$\mathcal{S}$ is the set of subjects defined in Definition 1;*
- *$\mathcal{R}$ is the set of relations linking any pair of subjects;*
- *$\mathcal{H}^{\mathcal{S}}_{\mathcal{R}}$ is the hierarchical structure of $\mathcal{O}$ constructed by $\mathcal{S} \times \mathcal{R}$.* □

3.2 Document Features

Various representations have been studied to formally describe text documents. The lexicon-based representation is based on the statistic of occurring terms. Such a representation is easy to understand by users and systems. However, along with meaningful, representative features, some noisy terms are also extracted, caused by sense ambiguity of terms. To deal with this problem, pattern-based representation is studied, which uses frequent sequential patterns (phrases) to represent document contents [9]. The pattern-based representation is superior to lexicon-based, as the context of terms co-occurred in phrases is considered. However, the pattern-based presentation suffers from a limitation caused by the length of patterns. Though a long pattern is wealthy with information and so more discriminative, it usually has low frequency and as a result, becomes inapplicable. To overcome the problem, we represent the content of documents by a set of weighted closed frequent sequential patterns discovered by pattern mining techniques.

Definition 3. (FEATURES) *Given a document $d = \{t_1, t_2, \ldots, t_n\}$ as a sequential set of repeatable terms, the feature set, denoted as $\mathcal{F}(d)$, is a set of weighted phrase patterns, $\{\langle p, w(p) \rangle\}$, extracted from d that satisfies the following constraints:*

- $\forall p \in \mathcal{F}(d), p \subseteq d$.
- $\forall p_1, p_2 \in \mathcal{F}(d)(p_1 \neq p_2), p_1 \not\subset p_2 \wedge p_2 \not\subset p_1$.
- $\forall p \in \mathcal{F}(d), w(p) \geqslant \vartheta$, *a threshold.* □

3.3 Initial Classification

The initial classification of d to $s_k \in \mathcal{S}$ is done through accessing a term-subject matrix created by the subjects and their labels. Adopting the features discovered previously, we use a feature-subject mapping approach to initially assign subject classes to the document.

Definition 4. (TERM-SUBJECT MATRIX) *Let $\mathcal{T}$ be the term space of $\mathcal{S}$, $\mathcal{T} = \{t \in \bigcup_{s\in\mathcal{S}} label(s)\}$, $\langle\mathcal{S},\mathcal{T}\rangle$ is the matrix coordinated by $\mathcal{T}$ and $\mathcal{S}$, where a mapping exists:*

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

and its reverse mapping also exists:

$$\mu^{-1} : \mathcal{S} \to 2^{\mathcal{T}}, \quad \mu^{-1}(s) = \{t \in \mathcal{T} | s \in \mu(t)\} \qquad \square$$

Adopting Definition 3 and 4, we can initially classify $d_i \in \mathcal{D}$ into a set of subjects using the following prediction:

$$\widehat{y}_i^k = I(s_k \in h \circ g \circ f(d_i)), i = 1, \ldots, m \tag{1}$$

where $I(z)$ is an indicator function that outputs 1 if z is true and zero, otherwise; $f(d) = \{p | \langle p, w(p)\rangle \in F(d)\}$; $g(\rho) = \{t \in \cup_{p\in\rho} p\}$; $h(\tau) = \{s \in \cup_{t\in\tau} \mu(t)\}$.

3.4 Generalised Classification

The initial classification process easily generates noisy subjects because of direct feature-subject mapping. Against the problem, we introduce a method to generalise the initial subjects to optimise the classification. We observed that in initial classification some subjects extracted from the ontology are overlapping in their semantic space. Thus, we can optimise the classification result by keeping only the dominating subjects and pruning away those being dominated. This can be done by investigating the semantic relations existing between 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 an broader semantic space than s_2 and thus, is more general. Vice versa, s_2 is more specific and focused than s_1. Hence, if some subjects are covered by a common ancestor, they can be replaced by the common ancestor without information loss. The common ancestor is unnecessary to be chosen from the initial classification result, as choosing an external common ancestor also satisfies the above rule. After generalising the initial classification result, we have a smaller set of subject classes, with no information lost but some focus. (The handling of focus problem is presented in next section.)

Definition 5. (GENERALISED CLASSIFICATION) *Given a document d and its initial classification result, a subject set denoted by $S^I(d)$, the generalised classification result, denoted as $S^G(d)$, is the set of subjects satisfying:*

1. $\forall s \in S^I(d), \exists s' \in S^G(d), s \neq s', s \in descendants(s')$.
2. $\forall s_1, s_2 \in S^G(d)(s_1 \neq s_2), s_1 \notin descendants(s_2) \wedge s_2 \notin descendants(s_1)$.

```
input : $d = \{t_1, t_2, \ldots, t_n\}$ where $n = |d|$, a threshold $\vartheta$.
output: The feature set $\mathcal{F}(d) = \{\langle p, w(p)\rangle\}$.
$P(d) = \emptyset, \mathcal{F}(d) = \emptyset, p = \emptyset$;
//Extracting sequential patterns;
for $(i = 1; i <= n; i++)$ do
    for $(j = i; j <= (n - i); j++)$ do
        $p = p \cup \{t_j\}$;
    end
    if $p \in P(d)$ then $w(p)++$ for $\langle p, w(p)\rangle \in \mathcal{F}(d)$else $P(d) = P(d) \cup \{p\}$,
    $\mathcal{F}(d) = \mathcal{F}(d) \cup \{\langle p, 1\rangle\}$;
end
//Filtering $\mathcal{F}(d)$ for closed, frequent patterns;
foreach $\langle p, w(p)\rangle \in \mathcal{F}(d)$ do
    if $w(p) < \vartheta$ then $\mathcal{F}(d) = \mathcal{F}(d) - \{\langle p, w(p)\rangle\}$else foreach $\langle p_k, w(p_k)\rangle \in \mathcal{F}(d)$ do
        if $p \subset p_k$ and $w(p) \leq w(p_k)$ then $\mathcal{F}(d) = \mathcal{F}(d) - \{\langle p, w(p)\rangle\}$
    end
end
return $\mathcal{F}(d)$.
```

Algorithm 1: Extracting Features from a Document

4 Implementation

In this section, we present the technical details for implementing the proposed approach of unsupervised multi-label text classification.

Algorithm 1 describes the process of extracting features to represent a document. The output is $\mathcal{F}(d)$, a set of closed frequent sequential patterns discovered from d. Adopting the prediction in Eq. (1), with $\mathcal{F}(d)$ the initial set of subjects, $S^I(d)$, can be assigned to classify d. Taking into account the weights of feature patterns, we can evaluate $t \in d$:

$$w(t) = \sum_{p \in \{p | t \in g \circ f(d), p \in f(d)\}} w(p)$$

All $s \in S^I(d)$ can then be re-evaluated for their likelihood of being assigning to d with consideration of term evaluation and term distribution in $s \in S^I(d)$. A prediction function can then be used to assess initial classification subjects for the second run of classification:

$$\widehat{y'}_i^{\kappa} = \mathcal{I}(\sum_{t \in \mu^{-1}(s_\kappa)} w(t) \times log(\frac{|S^I(d_i)|}{sf(t, S^I(d_i))}) \geqslant \theta), i = 1, \ldots, m \qquad (2)$$

where $\mathcal{I}(z \geqslant \theta)$ returns the value of z if $z \geqslant \theta$ is true and zero, otherwise; $\kappa = 1, ..., \mathcal{K}$ and $S^I(d) = \{s_1, \ldots, s_{\mathcal{K}}\}$ with $|S^I(d)| = \mathcal{K}$; θ is the threshold for filtering out noisy subjects. In experiments different values were tested for θ. The results revealed that setting θ as the top fifth z in $S^I(d_i)$, a variable rather than a static value, gave the best performance. (Refer to Section 6 for detail.)

In the generalisation phase, descendant subjects are replaced by their common ancestor subject. However, the common ancestor should not be too far away from the replaced descendants in the ontology structure. The focus will be significantly lost, otherwise. In implementation, we use only the lowest common

```
   input : $S_i = \{s_1, s_2, \ldots, s_j\}$ (subject classes assigned to $d_i$ after Eq. (2)), $\mathcal{O}$;
   output: $\mathcal{S}'_i = \{s_1, s_2, \ldots, s_k\}$ (subject classes generalised for optimising classification).
1  $\mathcal{S}'_i = \emptyset, S_{temp} = \emptyset, S_{redundant} = \emptyset$;
2  foreach $s \in S_i$ do
3      Extract $S(s)$ from $\mathcal{O}$ where $S(s) = \{s' | s' \in ancestor(s), \delta(s \mapsto s') \leq 3\}$; foreach
       $s_n \in S_i$ 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
           $\{\widehat{s} = \mathcal{LCA}(S(s) \cup S(s_n)), str(i, \widehat{s}) = str(i, s) + str(i, s_n)$; $S_{temp} = S_{temp} \cup \{\widehat{s}\}$;
           $S_{redundant} = S_{redundant} \cup \{s, s_n\}\}$
6      end
7      if $S_{temp} \neq \emptyset$ then $\{\mathcal{S}'_i = \mathcal{S}'_i \cup S_{temp}$; $S_i = S_i - S_{redundant}$; $S_{temp} = \emptyset$;
       $S_{redundant} = \emptyset\}$ else $\mathcal{S}'_i = \mathcal{S}'_i \cup \{s\}$
8  end
9  return $\mathcal{S}'_i$.
```

Algorithm 2: Generalising Subjects for Optimal Classification

ancestor (shortened by LCA) to replace its descendant subjects. The LCA is the common ancestor of a set of subjects, with the shortest distance to these subjects in the ontology structure. The LCA replaces descendant subjects with full information kept and minimised focus lost.

Algorithm 2 describes the process of generalising the initial subject classes to optimise classification. The function $str(i, s)$ describes the likelihood of assign s tod_i and returns the value of $\mathcal{I}(z \geqslant \theta)$ in Prediction function $\widehat{y'}_i^{\kappa}$ in Eq. (2). The function $\delta(s_1 \mapsto s_2)$ returns a positive real number indicating the distance between two subjects. Such a distance is measured by counting the number of edges travelled through from s_1 to s_2 in $\mathcal{H}_{\mathcal{R}}^{\mathcal{S}}$. The function $\mathcal{LCA}(S(s_1) \cup S(s_2))$ returns $\widehat{s}$, the LCA of s_1 and s_2. Note that $\delta(s_1 \mapsto s_2) \leq 3$, which restricts LCAs to three edges in distance. Subjects further than that in distance are too general; whereas using a highly-general subject for generalisation would severely jeopardise the focus of original subjects. (In the 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 LCA. The testing results revealed that as of three the distance was better.)

5 Evaluation

The experiments were performed, using a large corpus collected from the catalogue of a library using the LCSH for information organising. The title and content of each catalogue item were used to form the content of a document. The subject headings associated with the catalogue items were manually assigned by specialist librarians who were trained to specify subjects for documents without bias [4]. The documents and subjects provided an ideal ground truth in the experiments to evaluate the effectiveness of the proposed classification method. This objective evaluation methodology assured the solidity and reliability of the experimental evaluation.

The testing set was crawled from the online catalogue of library of the University of Melbourne[1]. General text pre-processing techniques, such as stopword

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

removal and word stemming (Porter stemming algorithm), were applied to the preparation of testing set for experiments. In the experiments, we used only documents containing at least 30 terms, resulted in 31,902 documents in the testing set. Documents shorter than that could hardly provided substantial frequent patterns for feature extraction, as revealed in the preliminary experiments.

Given that the LCSH ontology contains 394,070 subjects in our implementation, the problem actually became a K-class text classification problem where $K = |\mathcal{S}| = 394,070$, a very large number. Hence, we chose two typical multi-class classification approaches, *Rocchio* and kNN, as the baseline models in the experiments.

The performance of experimental models was measured by precision and recall, the modern evaluation methods in information retrieval and organising. In terms of text classification, precision was to measure the ability of a method to assign a document with only focusing subjects, and recall the ability to assign a document with all dealing subjects.

Taking into account $K = |\mathcal{S}| = 394070$, in respect with the testing document set and the ground truth featured by the LCSH, the classification performance was evaluated by:

$$precision = \frac{|\mathcal{FT}(S_{tgt}) \cap \mathcal{FT}(S_{grt})|}{|\mathcal{FT}(S_{tgt})|} \text{ and } recall = \frac{|\mathcal{FT}(S_{tgt}) \cap \mathcal{FT}(S_{grt})|}{|\mathcal{FT}(S_{grt})|}$$

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

F_1 Measure as another common method used in information organising systems was also employed in evaluation. We used *micro*-F_1, which evaluated each document's classification result first and then averaged the results for the final F_1 value. Greater F_1 values indicate better performance.

6 Results and Discussions

Naming our proposed unsupervised classification approach as the UTC model, the experiments were to compare the effectiveness performance of the UTC model to the baselines, Rocchio and kNN models. Their effectiveness performances are depicted in Fig. 1 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 classification meets the respective accuracy rate. As shown in the figure, the effectiveness rates were 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 Table 1.

Table 1. Effectiveness Performance on Average

	Precision	Recall	F-Measure
UTC	0.158	0.135	0.125
Rocchio	0.020	0.290	0.020
kNN	0.021	0.054	0.016

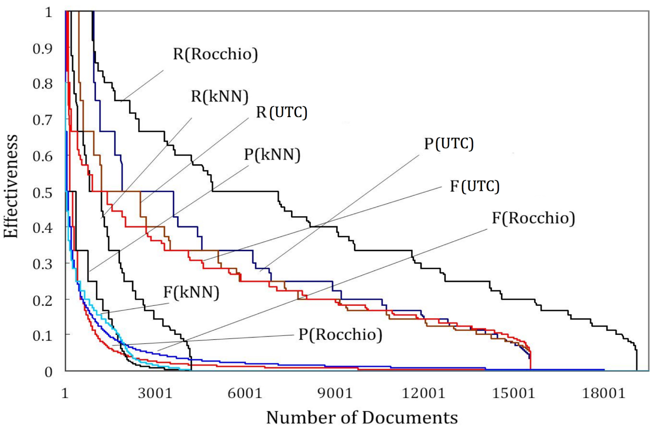

Fig. 1. Effectiveness Performance Results

F_1 Measure equally considers both precision and recall. Thus the F_1 Measure results can be deemed as an overall effectiveness performance. The average F_1 Measure result shown in Table 1 reveals that the UTC model has achieved a much better overall performance (0.125) than other two models (0.020 and 0.016). Such a performance is also confirmed by the detailed results depicted in Fig. 1 - the $F(UTC)$ line is located at much higher bound level compared to the $F(Rocchio)$ and $F(kNN)$ lines.

Precision measures how accurate the classification is. In terms of this, the UTC model once again has outperformed the baseline models. The average precision results shown in Table 1 demonstrates the achievement (UTC 0.158 vs. Rocchio 0.020 and *k*NN 0.021). The precision results depicted in Fig. 1 illustrate the same conclusion; the $P(UTC)$ outperformed others.

Recall measures the performance of classification by considering all dealing classes. The recall performance in the experiments shows a slightly different result, compared with those from F_1 Measure and precision performance. The *Rocchio* model achieved the best recall performance (0.290 on average), compared to that of the UTC model (0.135) and the *k*NN model (0.054). The result is also illustrated in Fig. 1, where $R(UTC)$ lies in the middle of $R(Rocchio)$ and $R(kNN)$.

There was a gap between the recall performance of the UTC and the *Rocchio* models. From the observation of recall results, we found that the classes assigned by the *Rocchio* model were usually a large set of subjects (935 on average), whereas the UTC model assigned documents with a reasonable number of subjects (16 on average) and the *k*NN results had an average size of 106. Due to the natural of recall measurement, more feature term would be cover if the subject size became larger. As a result, the *Rocchio* classification with the largest size achieved the best recall performance. The subject sets assigned by the *k*NN model had larger size than those assigned by the UTC. However, when expanding the classification by neighbours, a large deal of nosey data was also brought into the neighbourhood - the average number of neighbours arisen was 336. This was caused by the very large set and short length of documents in consideration.

As a result, the classification became inaccurate though only the documents with the top cosine values were chosen to expand and only the subjects with the top similarity values were chosen to classify a document.

Table 2. Performance Comparison for Finding the LCA

	Precision	Recall	F-Measure
Level = 3	0.158	0.135	0.125
Level = 5	0.154	0.112	0.111

Different number of levels were tested in sensitivity study for choosing a right number of levels to find the lowest common ancestor when generalising subjects for optimal classification. Table 2 displays the testing results for finding such a right level number. In the same experimental environment, if tracing three levels to find a LCA the UTC model's overall performance including F_1 Measure, precision, and recall was better than that of tracing five levels. In addition, tracing three levels only would give us better complexity. Therefore, we chose three levels to restrict the extent of finding CLAs.

7 Conclusions

Text classification has been widely exploited to improve the performance in information retrieval, information organising, text categorisation, and knowledge engineering. Traditionally, text classification relies on the quality of target categorises and the accuracy of classifiers learned from training samples. Sometimes qualified training samples may be unavailable; the set of categories used for classification may be with inadequate topic coverage. Sometimes documents may be classified into noisy classes because of large dimension of categories. Aiming to deal with these problems, in this paper we have introduced an unsupervised multi-label text classification method. Using a world ontology built from the LCSH, the method consists of three modules; closed frequent sequential pattern mining for feature extraction; extracting subjects from the ontology for initial classification; and generalising subjects for optimal classification. The method has been promisingly evaluated by compared with typical text classification methods, using a large real-world corpus, based on the ground truth encoded by human experts.

References

1. Ron Bekkerman and Matan Gavish. 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*, KDD '11, pages 231–239, 2011.
2. Deng Cai, Chiyuan Zhang, and Xiaofei He. Unsupervised feature selection for multi-cluster data. In *Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining*, KDD '10, pages 333–342, 2010.
3. Fabrice Camous, Stephen Blott, and Alan F. Smeaton. Ontology-based medline document classification. In *Proceedings of the 1st international conference on Bioinformatics research and development*, BIRD'07, pages 439–452, 2007.

4. Lois Mai Chan. *Library of Congress Subject Headings: Principle and Application*. Libraries Unlimited, 2005.
5. Evgeniy Gabrilovich and Shaul Markovitch. Feature generation for text categorization using world knowledge. In *Proceedings of The 19th International Joint Conference for Artificial Intelligence*, pages 1048–1053, 2005.
6. Michael Edward Houle and Nizar Grira. A correlation-based model for unsupervised feature selection. In *Proceedings of the 16th ACM conference on Conference on information and knowledge management*, CIKM '07, pages 897–900, 2007.
7. Xiaohua Hu, Xiaodan Zhang, Caimei Lu, E. K. Park, and Xiaohua Zhou. Exploiting wikipedia as external knowledge for document clustering. In *KDD '09: Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 389–396, 2009.
8. Ioannis Katakis, Grigorios Tsoumakas, and Ioannis Vlahavas. Multilabel text classification for automated tag suggestion. In *In: Proceedings of the ECML/PKDD-08 Workshop on Discovery Challenge*, 2008.
9. Yuefeng Li, Abdulmohsen Algarni, and Ning Zhong. Mining positive and negative patterns for relevance feature discovery. In *Proceedings of 16th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 753–762, 2010.
10. Hassan H. Malik and John R. Kender. Classifying high-dimensional text and web data using very short patterns. In *Proceedings of the 2008 8th IEEE International Conference on Data Mining, ICDM '08*, pages 923–928, 2008.
11. Leonardo Rocha, Fernando Mourão, Adriano Pereira, Marcos André Gonçalves, and Wagner Meira, Jr. Exploiting temporal contexts in text classification. In *Proceeding of the 17th ACM conference on Information and knowledge management*, CIKM '08, pages 243–252, 2008.
12. Xiaohui Tao, Yuefeng Li, and Ning Zhong. A personalized ontology model for web information gathering. *IEEE Transactions on Knowledge and Data Engineering, IEEE computer Society Digital Library*, 23(4):496–511, 2011.
13. Pu Wang and Carlotta Domeniconi. Building semantic kernels for text classification using wikipedia. In *KDD '08: Proceeding of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 713–721, 2008.
14. Yulan Yan, Naoaki Okazaki, Yutaka Matsuo, Zhenglu Yang, and Mitsuru Ishizuka. 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 Processing of the AFNLP: Volume 2*, ACL '09, pages 1021–1029, 2009.
15. Bishan Yang, Jian-Tao Sun, Tengjiao Wang, and Zheng Chen. Effective multi-label active learning for text classification. In *KDD '09: Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 917–926, 2009.
16. Tianbao Yang, Rong Jin, Anil K. Jain, Yang Zhou, and Wei Tong. Unsupervised transfer classification: application to text categorization. In *Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining*, KDD '10, pages 1159–1168, 2010.