Investigations of Term Expansion on Text Mining Techniques
Abstract

Recent advances in computer and network technologies have contributed significantly to global connectivity and stimulated the amount of online textual document to grow extremely rapidly. The rapid accumulation of textual documents on the Web or within an organization requires effective document management techniques, covering from information retrieval, information filtering and text mining. The word mismatch problem represents a challenging issue to be addressed by the document management research. Word mismatch has been extensively investigated in information retrieval (IR) research by the use of term expansion (or specifically query expansion). However, a review of text mining literature suggests that the word mismatch problem has seldom been addressed by text mining techniques. Thus, this thesis aims at investigating the use of term expansion on some text mining techniques, specifically including text categorization, document clustering and event detection. Accordingly, we developed term expansion extensions to these three text mining techniques. The empirical evaluation results showed that term expansion increased the categorization effectiveness when the correlation coefficient feature selection was employed. With respect to document clustering, techniques extended with term expansion achieved comparable clustering effectiveness to existing techniques and showed its superiority in improving clustering specificity measure. Finally, the use of term expansion for supporting event detection has degraded the detection effectiveness as compared to the traditional event detection technique.

Keywords: Term Expansion, Term Association, Word Mismatch, Text Mining, Text Categorization, Document Clustering, Event Detection
摘要

近來電腦及網路科技的快速發展促成了全球網路的連結，也使得線上文件快速地成長及累積。這些在網路上或組織內所累積下來的文件可能含有許多組織競爭所需的知識，有效的文件管理(Document Management)技術(包括資訊檢索(Information Retrieval)、資訊過濾(Information Filtering)、文字探勘(Text Mining)等)可協助組織有效的運用這些文件。然而，文件管理研究面臨一項挑戰性的議題，即所謂的字詞使用差異(Word Mismatch)。目前字詞使用差異的研究主要是在資訊检索的研究領域，並以字詞擴展(Term Expansion)的技術來解決這個問題，然而，在文件探勘的文獻中，這個問題卻極少被處理與解決。因此，本論文旨在對文件探勘技術中字詞擴展之使用進行研究，並特別以文件分類(Text Categorization)、文件分群(Document Clustering)以及事件偵測(Event Detection)這三類文件探勘技術為研究對象，發展這三類技術所需之字詞擴展技術。根據實證評估的結果，當使用相關係數(Correlation Coefficient)作為特徵選擇(Feature Selection)方式時，字詞擴展技術增加了的文件分類之效能。在文件分群方面，使用字詞擴展之文件分群技術並未改善分群之效能，但在 Specificity 的衡量上，使用字詞擴展技術的結果普遍明顯地優於傳統文件分群技術。最後，使用字詞擴展來協助事件偵測則導致了偵測效果的降低。

關鍵字: 字詞擴展、字詞關聯、字詞使用差異、文件探勘、文件分類、文件分群、事件偵測
TABLE OF CONTENTS

CHAPTER 1  INTRODUCTION ................................................................. 1
  1.1 BACKGROUND .................................................................................. 1
  1.2 RESEARCH MOTIVATION AND OBJECTIVES ...................................... 1
  1.3 ORGANIZATION OF THE THESIS........................................................ 3

CHAPTER 2  LITERATURE REVIEW ...................................................... 5
  2.1 TERM ASSOCIATION CONSTRUCTION ............................................... 5
  2.2 TEXT CATEGORIZATION ..................................................................... 9
  2.3 DOCUMENT CLUSTERING .................................................................. 14
  2.4 EVENT DETECTION ............................................................................ 16

CHAPTER 3  TEXT MINING TECHNIQUES WITH TERM EXPANSION ....... 20
  3.1 TERM ASSOCIATION CONSTRUCTION PROCESS ................................. 20
  3.2 TEXT CATEGORIZATION WITH TERM EXPANSION ............................. 22
  3.3 DOCUMENT CLUSTERING WITH TERM EXPANSION .......................... 24
  3.4 EVENT DETECTION WITH TERM EXPANSION .................................... 26

CHAPTER 4  EMPIRICAL EVALUATION FOR TEXT CATEGORIZATION WITH TERM EXPANSION ................................................................. 29
  4.1 EVALUATION DESIGN ........................................................................ 29
    4.1.1 Data collection ............................................................................. 29
    4.1.2 Evaluation Criteria ...................................................................... 31
    4.1.3 Evaluation Procedure .................................................................. 31
  4.2 EVALUATION RESULT ......................................................................... 32
    4.2.1 Effects of Number of Features on Categorization Effectiveness ........ 32
    4.2.2 Effects of Induction Method, Feature Selection Method and Representation Scheme on Categorization Effectiveness ............................................. 36
    4.2.3 Comparative Evaluation of Text Categorization Techniques ............ 38

CHAPTER 5  EMPIRICAL EVALUATION FOR TERM EXPANSION EMBEDDED DOCUMENT CLUSTERING ................................................................. 42
  5.1 EVALUATION DESIGN ........................................................................ 42
    5.1.1 Evaluation Criteria ...................................................................... 42
    5.1.2 Evaluation Procedure .................................................................. 44
  5.2 EVALUATION RESULT ......................................................................... 45
    5.2.1 Comparative Evaluation of Document Clustering Techniques .......... 45

CHAPTER 6  EMPIRICAL EVALUATION FOR EVENT DETECTION WITH
TERM EXPANSION................................................................................................................. 50

6.1 Evaluation Design........................................................................................................ 50
   6.1.1 Data Collection .................................................................................................. 50
   6.1.2 Evaluation Criteria ......................................................................................... 51
   6.1.3 Performance Benchmark .............................................................................. 51
   6.1.4 Evaluation Procedure ................................................................................... 52

6.2 Evaluation Result......................................................................................................... 52
   6.2.1 Parameter Tuning ........................................................................................... 52
   6.2.2 Comparative Evaluation of Event Detection Techniques .............................. 56

CHAPTER 7 CONCLUSION AND FUTURE RESEARCH DIRECTIONS ...... 57

REFERENCES ..................................................................................................................... 59
LIST OF FIGURES

FIGURE 2.1. EXAMPLE OF TERM ASSOCIATIONS ............................................................. 7
FIGURE 2.2. NETWORK OF ASSOCIATED TERMS [WBO00]............................................. 9
FIGURE 2.3. GENERAL PROCESS OF TEXT CATEGORIZATION PATTERN LEARNING .... 10
FIGURE 2.4. GENERAL PROCESS OF DOCUMENT CLUSTERING................................... 14
FIGURE 2.5. GENERAL PROCESS OF INCR................................................................. 18
FIGURE 3.1. PROCESS OF TERM ASSOCIATION CONSTRUCTION ......................... 21
FIGURE 3.2. LEARNING TEXT CATEGORIZATION PATTERNS WITH TERM EXPANSION .... 22
FIGURE 3.3. GENERAL PROCESS OF DOCUMENT CLUSTERING WITH TERM EXPANSION 25
FIGURE 3.4. GENERAL PROCESS OF INCR WITH TERM EXPANSION.......................... 27
FIGURE 4.1. EXAMPLE OF TERM ASSOCIATIONS CONSTRUCTED....................... 30
FIGURE 4.2. PRECISION RATES OF TRADITIONAL TEXT CATEGORIZATION TECHNIQUES (USING BINARY REPRESENTATION SCHEME AND C4.5 AS INDUCTION METHOD) ................................................................................................ 33
FIGURE 4.3. RECALL RATES OF TRADITIONAL TEXT CATEGORIZATION TECHNIQUES (USING BINARY REPRESENTATION SCHEME AND C4.5 AS INDUCTION METHOD) ................................................................................................ 33
FIGURE 4.4. PRECISION RATES OF TRADITIONAL TEXT CATEGORIZATION TECHNIQUES (USING BINARY REPRESENTATION SCHEME AND CN2 AS INDUCTION METHOD) .......................................................................................... 34
FIGURE 4.5. RECALL RATES OF TRADITIONAL TEXT CATEGORIZATION TECHNIQUES (USING BINARY REPRESENTATION SCHEME AND CN2 AS INDUCTION METHOD) .......................................................................................... 34
FIGURE 4.6. PRECISION RATES OF TEXT CATEGORIZATION WITH TERM EXPANSION (USING TF REPRESENTATION SCHEME AND CN2 AS INDUCTION METHOD) .................................................................................. 35
FIGURE 4.7. RECALL RATES OF TEXT CATEGORIZATION WITH TERM EXPANSION (USING TF REPRESENTATION SCHEME AND CN2 AS INDUCTION METHOD)...... 36

FIGURE 4.8. PRECISION RATES OF INDIVIDUAL CATEGORIES (USING CORRELATION COEFFICIENT SELECTION METHOD, BINARY REPRESENTATION SCHEME, AND C4.5 AS INDUCTION METHOD)......................................................... 40

FIGURE 4.9. RECALL RATES OF INDIVIDUAL CATEGORIES (USING CORRELATION COEFFICIENT SELECTION METHOD, BINARY REPRESENTATION SCHEME, AND C4.5 AS INDUCTION METHOD)......................................................... 41

FIGURE 5.1. CLUSTER RECALL RATES OF TRADITIONAL DOCUMENT CLUSTERING TECHNIQUE (USING TF SELECTION METHOD) ........................................ 46

FIGURE 5.2. PURITY RATES OF TRADITIONAL DOCUMENT CLUSTERING TECHNIQUE (USING TF×IDF SELECTION METHOD) ................................................... 46

FIGURE 5.3. DIVERSITY RATES OF TRADITIONAL DOCUMENT CLUSTERING TECHNIQUE (USING TF×IDF SELECTION METHOD) ................................................... 47

FIGURE 5.4. CLUSTER PRECISION RATES OF TRADITIONAL DOCUMENT CLUSTERING TECHNIQUE (USING TF×IDF SELECTION METHOD)............................... 47

FIGURE 6.1. PARAMETER TUNING EXPERIMENTS FOR THE INCR TECHNIQUE.............. 53

FIGURE 6.2. PARAMETER TUNING EXPERIMENTS FOR THE INCR TECHNIQUE WITH UNRESTRICTED TERM EXPANSION METHOD ...................................................... 54

FIGURE 6.3. PARAMETER TUNING EXPERIMENTS FOR THE INCR TECHNIQUE WITH BOUNDED TERM EXPANSION METHOD ................................................................. 55

FIGURE 6.4. DET CURVES OF DIFFERENT EVENT DETECTION TECHNIQUES ............. 56
LIST OF TABLES

TABLE 2.1. SEMANTIC RELATIONS IN WORDNET ........................................................... 6
TABLE 4.1. SUMMARY OF DATA CORPUS ...................................................................... 30
TABLE 4.2. CATEGORIZATION EFFECTIVENESS (USING C4.5 AS INDUCTION METHOD). 37
TABLE 4.3. CATEGORIZATION EFFECTIVENESS (USING CN2 AS INDUCTION METHOD). 37
TABLE 5.1. CLUSTERING EFFECTIVENESS WITHOUT TERM EXPANSION ................. 48
TABLE 5.2. CLUSTERING EFFECTIVENESS WITH TERM EXPANSION............................ 48
TABLE 6.1. SOME EXAMPLES OF EVENT IN EACH TOPICS ......................................... 51
TABLE 6.2. SUMMARY OF DATA CORPUS FOR EVENT DETECTION EVALUATION .......... 51
Chapter 1
Introduction

1.1 Background
Recent advances in computer and network technologies have contributed significantly to global connectivity and stimulated the amount of online information in the format of free-format text to grow extremely rapidly. For example, the size of Alexa website (www.alexa.com) is over 2 billion documents and the LexisNexis database contains more than 1 billion documents (www.lexisnexis.com). It is also commonly observed that many organizations create and maintain huge volumes of textual information in the emerging knowledge economy and electronic commerce environment. As a result, there is a pressing need to support efficient and effective document management, including information retrieval, information filtering, text mining, etc.

In document management research, a challenging research issue has been identified, i.e., the word mismatch problem. It refers to the phenomenon that people use different words to describe the same concepts. Imaginably, one author might use the term “merger” to describe the concept of “business merger” in one document, while another might use the term “acquisition” in another document. These two terms are conceptually identical or highly similar for human being. According to [FLG87], two people use the same terms to describe a concept less than 20% of the time. Therefore, without properly addressing this word mismatch problem, highly related documents might be regarded as being dissimilar due to differences in their feature spaces.

1.2 Research Motivation and Objectives
Word mismatch has been extensively investigated in information retrieval (IR)
research by the use of term expansion (or specifically query expansion in the context of IR) [X97, XC96]. Given a thesaurus, additional terms that are semantically similar to or statistically associated with the initial query terms are added into the user query. Subsequently, the expanded query is used for information retrieval. With query expansion, the probability of retrieving relevant documents is higher in the presence of word mismatch. Evidently, the core component in query expansion is a thesaurus, manually or automated constructed from a set of document corpus. The manual-based thesaurus approach suffers from being time-consuming and knowledge intensive, while the automatic thesaurus approach automatically constructs a thesaurus (or term associations) from a collection of documents, thus more efficient and less knowledge intensive. According to empirical evaluation results, information retrieval supported by query expansion improves the retrieval effectiveness measured by recall and precision rates [XC96, X97].

Due to advances of data mining techniques, research on text mining is emerging and gaining much research attention. Text mining refers to the discovery of interesting and useful patterns from a vast amount of documents for critical decision support or improving effectiveness of document management [WPS02]. Generally speaking, text mining encompasses such techniques as text categorization, document clustering, event detection, event tracking, information extraction (from unstructured or semi-structured documents), etc. These text mining techniques also suffer from the word mismatch problem. However, a review of text mining literature suggests that the word mismatch problem has seldom been addressed by text mining research. Therefore, the objective of this thesis is to investigate the use of term expansion on several text mining techniques, specifically including text categorization, document clustering, and event detection. Text categorization refers to the assignment of textual
documents, on the basis of their contents, to one or more predefined categories [DPH98, YC94]. Benefits of a text categorization include the ability to quickly locate a document without having to remember the exact keywords contained in that document, and the ability to easily browse a set of related documents [ABS99].

*Document clustering*, an unsupervised learning method, groups similar documents into separate clusters. The documents included in the respective clusters exhibit maximal intra-cluster similarity and minimal inter-cluster similarity. On the other hand, the objective of event detection is to identify stories in several continuous news streams that pertain to new or previously unidentified events [YPC98]. Event detection can assist organizations’ environmental scanning, the first link in the chain of perceptions and actions that permit an organization to adapt to its environment and subsequently to develop effective responses to secure or improve their position in the future [H81, JL92, C99].

Specifically, in this thesis, we will develop appropriate term expansion extensions to existing text categorization, document clustering and event detection techniques. Using traditional text mining techniques as performance benchmarks, we will empirically evaluate the applicability of term expansion extensions to different text mining techniques.

### 1.3 Organization of the Thesis

This chapter presented the motivation and objective of this thesis study. The rest of the thesis is organized as follows. In Chapter 2, the literatures relevant to this thesis will be reviewed, including existing term association construction methods, text categorization, document clustering, and event detection techniques. The development of the target text mining techniques with term expansion extensions will be depicted
in Chapter 3. An empirical evaluation of text categorization with term expansion using a set of news stories will be conducted and summarized in Chapter 4. In Chapter 5, the empirical evaluation design and results of document clustering with term expansion are provided. In Chapter 6, we will discuss the empirical evaluation of event detection techniques supported by term expansion. Finally, the contributions of this thesis study as well as some future research work will be concluded in Chapter 7.
Chapter 2
Literature Review

In this chapter, literature related to this research will be reviewed, including term association construction in Section 2.1, text categorization in Section 2.2, document clustering in Section 2.3, and event detection in Section 2.4.

2.1 Term Association Construction

Given a thesaurus, additional terms that are semantically similar to or statistically associated with the initial query terms are added into the user query. Subsequently, the expanded query is used for information retrieval.

Two major approaches to construct a thesaurus include manual-based or association-based. Manual thesauri are compiled by linguists and domain experts. A well-known manual thesaurus is WordNet [M95]. In WordNet, a variety of semantic relations (see Table 2.1 for details) can be defined between words. Besides this general-purpose thesaurus, several domain-specific thesauri have been constructed, including UMLS (Unified Medical Language System) [N01] and MeSH (Medical Subject Headings) [M80]. During query expansion, synonymous (or other semantic relations) words of initial query terms are selected and an appropriate weight is assigned to each term. As mentioned, the expanded query is then used for retrieving relevant documents.
Table 2.1. Semantic Relations in WordNet

<table>
<thead>
<tr>
<th>Semantic Relation</th>
<th>Syntactic Category</th>
<th>Examples</th>
</tr>
</thead>
<tbody>
<tr>
<td>Synonymy (similar)</td>
<td>N, V, Aj, Av</td>
<td>pipe, tube rise, ascend sad, unhappy rapidly, speedily</td>
</tr>
<tr>
<td>Antonymy (opposite)</td>
<td>Aj, Av, (N, V)</td>
<td>wet, dry powerful, powerless friendly, unfriendly rapidly, slowly</td>
</tr>
<tr>
<td>Hyponymy (subordinate)</td>
<td>N</td>
<td>sugar maple, maple tree, plant</td>
</tr>
<tr>
<td>Meronymy (part)</td>
<td>N</td>
<td>brim, hat gin, martini</td>
</tr>
<tr>
<td>Troponymy (manner)</td>
<td>V</td>
<td>march, walk whisper, speak</td>
</tr>
<tr>
<td>Entailment</td>
<td>V</td>
<td>drive, ride divorce, marry</td>
</tr>
</tbody>
</table>

Note: N = Nouns, Aj = Adjectives, V = Verbs, Av = Adverbs

The manual-based thesaurus creation approach suffers from being time-consuming and knowledge intensive. In response, the association-based thesaurus construction approach that automatically constructs a thesaurus (or term associations) from a collection of documents has been proposed. A term association is composed of a pair of terms and the association strength between them. Figure 2.1 shows an example of a set of term associations. For example, the terms “data mining” and “knowledge discovery” are correlative and the strength of the relation is 0.85.
The underlying hypothesis for constructing term associations is the so-called association hypothesis: related terms in a collection of documents tend to co-occur in the documents of the corpus \([v79]\). To construct term associations, each term \(k_i\) in the corpus is associated with a vector \(\vec{k}_i\), given by

\[
\vec{k}_i = (w_{i,1}, w_{i,2}, \ldots, w_{i,N})
\]

where \(N\) is the number of documents in the collection, and

\(w_{i,j}\) is the weight of term \(k_i\) in document \(d_j\).

The relationship between two terms \(k_u\) and \(k_v\) is then computed as a correlation factor \(C_{u,v}\) which is then normalized as \(S_{u,v}\):

\[
C_{u,v} = \vec{k}_u \cdot \vec{k}_v = \sum_{i=1}^{N} w_{u,i} \times w_{v,i}
\]

\[
S_{u,v} = \frac{C_{u,v}}{\sqrt{\sum_{i=1}^{N} w_{u,i}^2 \times \sum_{i=1}^{N} w_{v,i}^2}}
\]
Different schemes for assigning term weight \( w_{i,j} \) have been proposed. The standard version of the TF scheme assigns the frequency (\( \text{tf}_{i,j} \)) of term \( k_i \) in document \( d_j \) as the weight \( w_{i,j} \) [AF77]. In the TF\times IDF scheme [SB90],

\[
    w_{i,j} = \frac{1 + \log_2 \text{tf}_{i,j}}{\|d\|} \times \text{log}_2 \left( \frac{N}{n_i} \right)
\]

(2.4)

where \( w_{i,j} \) is the weight of term \( k_i \) in document \( j \),

\( \text{tf}_{i,j} \) is the within-document term frequency (TF),

\( N \) is the number of documents in the collection,

\( n_i \) is the number of documents where term \( k_i \) occurs,

\( \log_2(N/n_i) \) is the inverse document frequency (IDF), and

\[
    \|d\| = \sqrt{\sum_{i} w_{i,j}^2}
\]

is the 2-norm of vector \( \vec{d} \).

The scheme proposed by Qiu and Frei [QF93] uses inverse term frequency (ITF) rather than IDF when determining a term weight:

\[
    w_{i,j} = \left( 0.5 + 0.5 \frac{\text{tf}_{i,j}}{\max_{j} \text{tf}_{i,j}} \right) \times \log \frac{t}{t_j} \frac{1}{\sqrt{\sum_{i=1}^{N} \left( 0.5 + 0.5 \frac{\text{tf}_{i,j}}{\max_{j} \text{tf}_{i,j}} \right) \left( \log \frac{t}{t_j} \right)^2}}
\]

(2.5)

where \( t \) is the number of terms in the collection,

\( t_j \) is the number of distinct terms in document \( j \), and

\( \log(t/t_j) \) is the inverse term frequency (ITF).
On the other hand, mining term association rules [WBO00] represents another method for constructing term associations, using the association rule analysis technique [AIS93, AS94]. According to Wei et al. (2000), each document is treated as a transaction and terms in the document are as the items in the transaction [WBO00]. Given the minimum support and confidence thresholds, association rule analysis technique generates a set of association rules each of which relates only two terms (e.g. treaty \( \rightarrow \) war, support = 0.01219, confidence=0.46269). The associations between terms can be represented as a network of associated terms [WBO00] as shown in Figure 2.2. The mining term association rule technique is more flexible than previous techniques based on term co-occurrence frequency since it takes into account not only the co-occurrence frequency but also the confidence and direction of the association rules [WBO00].

Figure 2.2. Network of Associated Terms [WBO00]

### 2.2 Text Categorization

Text categorization refers to the assignment of textual documents, on the basis of their contents, to one or more predefined categories [DPH98, YC94]. Traditionally, text categorization was performed manually by domain experts. However, the manual
process is very time-consuming and costly. To overcome the inefficiency of the manual text categorization process, the learning-based text categorization has been developed. The learning-based text categorization first discovers the text categorization patterns from a set of pre-classified documents known as training documents. Once the text categorization patterns are discovered, new documents can be categorized accordingly.

A challenging research issue in learning-based text categorization is the development of a learning algorithm for automatically inducing the text categorization patterns from the pre-classified training examples. In general, text categorization pattern learning consists of three main phases: feature extraction and selection, document representation, and induction [ADW94, WHD02], as shown in Figure 2.3.
In the feature extraction and selection phase, one or multiple sets of features (i.e., a universal dictionary or several local dictionaries) are created from and therefore used to represent training documents. A universal dictionary is created for all categories, whereas a local dictionary is pertinent to a particular category. Feature extraction commences with the parsing of each training document to produce a list of nouns or noun phrases referred to as features. These features usually exclude a set of pre-specified stop-words, numeric numbers and proper names. Following extraction is feature selection that trims the size of the feature set(s) extracted. Feature selection is crucial for learning efficiency and effectiveness because it condenses the feature set and, at the same time, reduces biases in the original feature set(s) [DPH98]. Several feature selection methods have been proposed, including within-document frequency (TF), document frequency×inverse document frequency (TF×IDF), correlation coefficient, $\chi^2$ metric, and mutual information [DPH98, LH98, LR94, NGL97, SHP95]. In the TF selection method, normalized within-document frequencies are often used so that training documents of different lengths are normalized to contribute equally during training. On the other hand, TF×IDF selection metric is defined as

$$TF \times IDF(k_i, C) =freq_{(k_i, c)} \times \left[ \log \frac{m}{df_{k_i}} + 1 \right]$$  \hspace{1cm} (2.6)

where $freq_{(k_i, c)}$ is the frequency of term $k_i$ that appears in the category $C$,

$df_{k_i}$ is the number of categories that contains term $k_i$, and

$m$ is the total number of categories.

The correlation coefficient $CC$ of a term $k_i$ relevant to a particular category $C$ is defined as:
\[ CC(k_i, C) = \frac{(n_{r+}n_{n-} - n_{r-}n_{n+})\sqrt{N}}{\sqrt{(n_{r+} + n_{r-})(n_{n+} + n_{n-})(n_{r+} + n_{n-})(n_{r-} + n_{n+})}} \] (2.7)

where \( n_{r+} \) is the number of documents in category \( C \) in which term \( k_i \) occurs,
\( n_{r-} \) is the number of documents in category \( C \) in which term \( k_i \) does not occur,
\( n_{n+} \) is the number of documents in the categories other than \( C \) in which term \( k_i \) occurs,
\( n_{n-} \) is the number of documents in the categories other than \( C \) in which \( k_i \) does not occur, and
\( N \) is the total number of documents.

The \( \chi^2 \) metric [SHP95] is the square of correlation coefficient:
\[ \chi^2(k_i, C) = [CC(k_i, C)]^2 \] (2.8)

Correlation coefficient selects exactly those terms that are highly indicative of membership in a category, whereas the \( \chi^2 \) metric will not only select from this set of terms but also those terms that are indicative of non-membership in the category.

The mutual information \( MI(k_i, C) \), between a feature \( k_i \) and a category \( C \) is defined as [DPH98, LR94]:
\[ MI(k_i, C) = \sum_{k_i \in \{0,1\}} \sum_{C \in \{0,1\}} P(k_i, C) \log \frac{P(k_i, C)}{P(k_i)P(C)} \] (2.9)

where \( P(k_i, C) \) is the probability that term \( k_i \) and the category \( C \) present together,
\( P(k_i) \) is the probability that term \( k_i \) appears, and
\( P(C) \) is the probability that the category \( C \) appears.

The top \( k \) features with the highest selection metric score are then selected as features.
to represent each document. In the document representation phase, each training
document is represented using the features included in the dictionary or dictionaries
generated in the feature extraction and selection phase. A document is labeled to
indicate its category membership and includes a value for each feature from the
dictionary or dictionaries of choice. The category membership of a document
resembles the decision class in typical classification problems in machine learning;
thus, specifying the particular category to which the document belongs. According to
the dictionary or dictionaries selected, a document is jointly described by multiple
features, each of which can assume a boolean or numerical value to indicate the
feature presence or occurrence frequency in the document. Several document
representation methods have also been proposed, including binary, TF, IDF\(^1\) and
TF×IDF [YC94].

In the induction phase, text categorization patterns that distinguish categories from
one another are automatically learned from a set of training documents. Common
strategies for automatic learning of text categorization patterns can be classified into
several types, including decision tree induction [WAD99]; decision rule induction
[ADW94, CS96]; k-nearest neighbor classification [IT95, LC96, MLW92, Y94];
neural network [WPW95, NGL97]; Bayesian networks [ABS99, BM98, LC96, LR94,
MN98]; and regression approach [YC94]. A detailed analysis and empirical
evaluation of common techniques for text categorization pattern induction can be

\[^{1}\text{IDF for the feature } k_i \text{ is:}\]

\[a_{i,j} = IDF_i = \log\left(\frac{N}{n_i}\right) + 1\]

where \(N\) is the number of documents in the entire collection, and
\(n_i\) is the number of documents in which the feature \(k_i\) is present.
found in [YL99].

2.3 Document Clustering

Document clustering, an unsupervised learning method, groups similar documents into separate clusters. The documents included in the respective clusters exhibit maximal intra-cluster similarity and minimal inter-cluster similarity. The general process of document clustering is similar to that of text categorization, broadly consisting of feature extraction and selection, document representation, and clustering, as shown in Figure 2.4. However, document clustering requires no pre-specified categories created manually or otherwise, because of its unsupervised learning orientation.

In document clustering, the feature extraction and selection phase is highly similar if
not identical to that of text categorization. Common feature selection methods for document clustering include TF, TF×IDF, and their hybrids. The top $k$ features with the highest feature selection metric score are selected as features to represent each document. The document representation in document clustering also closely resembles that of text categorization. Examples of common document representation methods include binary, TF, and TF×IDF.

In the clustering phase, documents are segmented into clusters, based on the selected features and their respective values for each document. Common approaches for document clustering include partitioning-based (e.g., [LA99]), hierarchical (e.g., [V86, EW86, RC99]), and Kohonen neural network approach (e.g., [RC99]). A partitioning-based approach partitions a set of documents into multiple non-overlapping clusters. Common partitioning-based algorithms include K-means [A73] and PAM (Partitioning Around Medoids) [KR90], CLARA (Clustering LARge Applications) [KR90], CLARAN (Clustering Large Applications based on RANdomized Search) [NH94], and genetic-algorithm-based clustering method [EM97]. A hierarchical clustering approach builds a binary clustering hierarchy whose leaf nodes denote documents to be clustered. A representative hierarchical clustering algorithm is the hierarchical agglomerative clustering (HAC) method, which starts with as many clusters as there are documents (i.e., a cluster contains one document only) [V86]. The two clusters bearing the most similarity are merged to form a cluster. The described merging process continues until a hierarchy of clusters is built, with a single cluster at the top of the hierarchy that contains all the documents. A Kohonen neural network, also known as a self-organizing map (SOM), is an unsupervised two-layer neural network [K89, K95]. With a Kohonen neural network or SOM, each input node corresponds to a coordinate axis in the input attribute vector space. Each
output node represents a node in a two-dimensional grid. The network is fully connected; that is, each output node is connected to every input node with a connection weight. During the training phase, the documents to be clustered are fed into the network multiple times in order to train or adjust the connection weights in such a way that distribution of the output nodes represents that of the input objects.

### 2.4 Event Detection

The objective of event detection is to identify stories in several continuous news streams that pertain to new or previously unidentified events [YPC98]. Event detection is subdivided into two forms: retrospective detection and online detection [YPC98, YCB99, APL98]. The former entails the discovery of previously unidentified events in a chronologically ordered accumulation of documents (stories), and the latter strives to identify the onset of new events from live news feeds in real-time. Both forms of detection intentionally lack advanced knowledge of novel events, but do have access to unlabelled historical news stories for use as contrast sets.

Most of the proposed event detection algorithms, retrospective or online, were developed based on the document clustering approach. Yang et al. [YPC98, YCB99] implemented two clustering methods for event detection: GAC and INCR. GAC, operating in a strict retrospective detection setting, performs agglomerative clustering, producing hierarchically organized document clusters. GAC employed the conventional vector space model to represent documents and clusters. Each document is represented using a vector of weighted terms, based on the $\text{TF} \times \text{IDF}$ (within-document frequency $\times$ inverse document frequency) scheme:
\[ TF \times IDF(t, d) = \left( \frac{(1 + \log_2 tf(t, d)) \times idf(t)}{\|d\|} \right) \]  
\[(2.10)\]

where \( tf(t, d) \) is the within-document term frequency (TF) of term \( t \) in document \( d \),
\( idf(t) \) is the inverse document frequency (IDF) of term \( t = \log_2 \left( \frac{N}{n(t)} \right) \),
\( N \) is the number of the training documents used to compute the IDF,
\( N(t) \) is the number of training documents where \( t \) occurs, and
\[ \|d\| = \sqrt{\sum w(t, d)^2} \] is the 2-norm of vector \( \tilde{d} \).

For cluster representation, the normalized vector of documents in a cluster is summed and the \( k \) most significant terms called the prototype or centroid of the cluster are selected to represent the cluster [YPC98]. To improve the computation efficiency and to preserve the characteristics that events tend to appear in news bursts, GAC adopted a divide-and-conquer strategy that grows clusters iteratively in a bottom-up fashion. In each iteration, the current pool of clusters is divided according to their order in time into evenly sized buckets. Subsequently, group-average clustering\(^2\) is applied to each bucket locally, merging smaller clusters into larger ones. Periodically, the news stories within each of the top-level clusters are reclastered. Reclustering is useful when events straddle the initial temporal-bucket boundaries or when the bucketing causes undesirable groupings of stories about different events.

On the other hand, INCR whose process is shown in Figure 2.5, is designed for both retrospective and online detection [YPC98, YCB99]. It is a single-pass incremental

\(^2\) Group-average clustering maximizes the average similarity between document pairs in the resulting clusters by merging clusters in a greedy, bottom-up fashion.
clustering algorithm that produces nonhierarchical clusters incrementally. For retrospective detection, the TF×IDF scheme was adopted to represent documents or clusters. However, to deal with the problem of continuously incoming documents that might affect term weighting and vector normalization during online detection, the incremental IDF was employed by INCR:

$$\text{idf}(t, p) = \log_2 \left( \frac{N(p)}{n(t, p)} \right)$$ (2.11)

where $p$ is the current time point,

$N(p)$ is the number of documents accumulated up to the current point (including the retrospective corpus if used), and

$n(t, p)$ is the document frequency of term $t$ at time $p$.

---

Figure 2.5. General Process of INCR
Moreover, INCR incorporated a time penalty when calculating the similarity between a document \( x \) and any cluster \( c \) in the past. The time penalty can be a uniformly weighted time window (i.e., a time window of \( m \) documents before \( x \) is imposed) or a linear decaying-weight function (shown as below).

\[
similarity(x, c) = \begin{cases} 
(1 - \frac{i}{m}) \times \text{similarity}(x, c) & \text{if } c \text{ has any member in the time window} \\
0 & \text{otherwise}
\end{cases}
\]  

(2.12)

where \( i \) is the number of documents between \( x \) and the most recent member document in \( c \), and \( m \) is the time window of documents before \( x \).

For retrospective detection, INCR sequentially processes news documents. A document is absorbed by the most similar cluster in the past if the similarity between the document and the cluster is larger than a pre-selected clustering threshold \( (t_c) \); otherwise, the document becomes the seed of a new cluster. For online detection, the novelty threshold \( (t_n) \) was introduced. If the maximal similarity between the current document and any cluster in the past is no less than \( t_n \), the document is flagged as containing an old event.
Chapter 3

Text Mining Techniques with Term Expansion

This chapter discusses the use of term expansion in supporting text mining techniques, specifically text categorization, document clustering, and event detection. We will first detail our term association construction process in Section 3.1. Text categorization supported by term expansion is depicted in Section 3.2, followed by document clustering with term expansion in Section 3.3. Finally, event detection with term expansion is elaborated in Section 3.4.

3.1 Term Association Construction Process

As shown in Figure 3.1, the term association construction process commences with extracting and selecting features from a collection of documents for establishing term associations. A set of nouns and noun phrases from this document corpus is first extracted. In this study, a rule-based part of speech tagger proposed by Brill [B92, B94] was adopted for syntactically tagging each word in the documents. For extracting noun phrases from syntactically tagged documents, a noun phrase parser proposed by Voutilainen [V93] was employed.
After feature extraction, feature selection is initiated to reduce the number of unnecessary features, a process that improves the efficiency and effectiveness of term association construction. Assume $DF$ is the number of documents that includes the particular term under discussion. Using $DF$ as the selection metric, we select those terms whose $DF$ is higher than a pre-specified threshold (e.g., 1% of the total number of documents for term association construction).

Using the set of features extracted and selected from the previous phase, the association strength between each pair of terms is then calculated. In this study, term $k_i$ in the corpus is associated a vector $\vec{k}_i = (w_{i,1}, w_{i,2}, ..., w_{i,N})$ where $N$ is the number of documents in the collection, and $w_{i,j}$ is the weight of term $k_i$ in document $d_j$. The weight $w_{i,j}$ is assigned based on the standard version of TF×IDF scheme [SB90] as shown in the formula (2.4), while the association strength between two terms is computed as a correlation factor $S_{u,v}$ as depicted in the formula (2.3).
3.2 Text Categorization with Term Expansion

The purpose of term expansion for supporting text categorization is to address the word mismatch problem potentially hampering the categorization effectiveness by extending the concept space of documents. The general process for text categorization (specifically, learning text categorization patterns) with term expansion is shown in Figure 3.2. This process is similar to the traditional text categorization pattern learning process as shown in Figure 2.3, with an additional phase, i.e., term expansion.

During the text categorization pattern learning process, the feature extraction phase is first initiated. As with the term association construction process described previously,
the rule-based part of speech tagger [B92, B94] and the noun phrase parser [V93] were employed for extracting noun phrases as features for each training document. Subsequently, during the term expansion phase, each feature $k_i$ extracted from a document $d_j$ will be expanded with additional $m$ terms that are associated with $k_i$. To select the $m$ terms to expand for $k_i$, the weight of each term associated with $k_i$ needs to be determined. Assume $w_{i,j}$ be the within document frequency of $k_i$ in $d_j$, and $s_{i,p}$ be the association strength between terms $k_i$ and $k_p$. Thus, in the document $d_j$, the weight of $k_p$ expanded from $k_i$ (i.e., $w_{i,j}$) is $w_{i,j} \times s_{i,p}$. Accordingly, the top $m$ terms with the highest weight are then selected as the expanded features for $k_i$ in $d_j$. If a term $k_p$ that is expanded from $k_i$ has already existed either in the original feature set or in the expanded feature set extended from other original features, the weight for $k_p$ is the larger weight attained.

In the feature selection phase, representative features will be selected from the feature sets (including original and expanded) of all training documents based on some feature selection metric. In this study, local dictionaries (i.e., a set of representative feature set for each category) were constructed using the correlation coefficient, chi-square ($\chi^2$) metric, or TF×IDF feature selection method. Afterwards, for each document category, all training documents are represented based on the local dictionary for the target category. We adopted the binary and TF schemes as alternative document representation methods in this study. Since the concept space of each document has been expanded by the term expansion phase, the binary scheme sets the value of the feature $f_i$ in the document $d_j$ as 1 if $f_i$ appears in the original or expanded feature set of $d_j$; otherwise, 0. On the other hand, the TF representation scheme takes the weight of the feature $f_i$ in the document $d_j$ (i.e., $w_{i,j}$) if $f_i$ is in the original or expanded feature set of $d_j$; otherwise 0. Finally, in the induction phase, two
classifiers, C4.5 (a decision tree induction technique) and CN2 (a decision rule induction technique), were adopted.

To classify an unseen document into appropriate category (or categories), the target document will go through the feature extraction and the term expansion phase. To predict whether the target document belongs to a particular category $C$, this document is represented with the dictionary created for $C$, using the binary or TF scheme. Subsequently, the text categorization pattern (e.g., a decision tree or a set of decision rules) pertaining to $C$ is reasoned. Since each category proceeds independently, the document may be assigned to more than one category or cannot be assigned to any category.

### 3.3 Document Clustering with Term Expansion

The general process for document clustering with term expansion is shown in Figure 3.3. This process is the same as the traditional document clustering process, as shown in Figure 2.4, but with an additional phase, i.e., term expansion.
During the document clustering pattern learning process, the feature extraction phase is first initiated. Specifically, the rule-based part of speech tagger [B92, B94] and the noun phrase parser [V93] were employed for extracting noun phrases as features for each document. Subsequently, during the term expansion phase, each feature $k_i$ extracted from a document $d_j$ will be expanded with additional $m$ terms that are associated with $k_i$. To select the $m$ terms to expand for $k_i$, the weight of each term associated with $k_i$ needs to be determined. Assume $w_{i,j}$ be the within document frequency of $k_i$ in $d_j$, and $s_{i,p}$ be the association strength between terms $k_i$ and $k_p$. Thus, in the document $d_j$ the weight of $k_p$ expanded from $k_i$ (i.e., $w_{i,j}$) is $w_{i,j} \times s_{i,p}$. Accordingly, the top $m$ terms with the highest weight are then selected as the expanded features for $k_i$ in $d_j$. If a term $k_p$ that is expanded from $k_i$ has already existed either in the original feature set or in the expanded feature set extended from other original features, the
weight for \( k_p \) is the larger weight attained.

In the feature selection phase, representative features will be selected from the feature sets (including original and expanded) of all training documents based on some feature selection metric. In this study, representative features were constructed using the TF or TF×IDF feature selection method. Afterwards, all documents are represented by the selected feature set. We adopted the binary, TF, and TF×IDF schemes as document representation methods in this study. Since the concept space of each document has been expanded by the term expansion phase, the binary scheme sets the value of the feature \( f_i \) in the document \( d_j \) as 1 if \( f_i \) appears in the original or expanded feature set of \( d_j \); otherwise, 0. On the other hand, the TF representation scheme takes the weight of the feature \( f_i \) in the document \( d_j \) (i.e., \( w_{i,j} \)) if \( f_i \) is in the original or expanded feature set of \( d_j \); otherwise 0. Conversely, the TF×IDF representation scheme take the TF×IDF value (the within document frequency of feature \( k_i \) in document \( d_j \times \) the inverse document frequency of feature \( k_i \) of feature \( f_i \) in the document \( d_j \) (i.e., \( w_{i,j} \)) if \( f_i \) is in the original or expanded feature set of \( d_j \); otherwise 0. Finally, in the clustering phase, CLARA (a partitioning-based clustering technique), was adopted.

### 3.4 Event Detection with Term Expansion

In traditional feature-based event detection techniques, whether a new news document contains a new event is determined by comparing the similarity of features between the new news document and past news documents. However, as mentioned, there exist vocabulary discrepancies between reporters even when they describe the same event. For example, some may use “merger” or “purchase” to describe a business merger event, while others may use “acquisition” for the same event. Such word
mismatch problem may impair the detection effectiveness. To address this word mismatch problem encountered in event detection, the concept space of the new news document as well as that of each past news document are extended by the use of term expansion technique, as shown in Figure 3.4. Once the terms in all documents involved in a detection process are expanded, a traditional event detection technique is applied to arrive at the detection decision for the new news document.

![Figure 3.4. General Process of INCR with Term Expansion](image)

Two expansion methods are proposed in this study to support event detection. The first method unrestrictedly expands the original terms in a news document (new or past). Thus, it is referred to as the *unrestricted expansion* method in this study. As
with the text categorization techniques with term expansion, when expanding a term \( k_i \) to \( k_p \) for a particular news document \( d_j \), the weight of \( k_p \) is determined as the association strength between \( k_i \) and \( k_p \) times the weight of \( k_i \) (i.e., the within document frequency of \( k_i \) in \( d_j \)). If \( k_p \) has already existed in the set of terms for \( d_j \), the weight of \( k_p \) in this set is updated if the newly derived weight is higher than the existing weight.

After the term expansion for the new news documents as well as for past news documents, the traditional event detection process is initiated. Specifically, we adopted the INCR technique for such detection process in this study.

The second expansion method restricts the expansion scope for a target document. Assume that the new news document be \( d_n \) and a past news document be \( d_o \). Let the set of original terms in \( d_n \) be \( T(d_n) \) and that in \( d_o \) be \( T(d_o) \). To extend the concept space of \( d_n \), the expanded terms include only those terms in \( T(d_o) - T(d_n) \). Likewise, the expansion of \( d_o \) introduces only those terms in \( T(d_n) - T(d_o) \). Since this method limits the expansion scope for one document by the original concept space of another document, it is referred to as the \textit{bounded expansion} method.
Chapter 4

Empirical Evaluation for Text Categorization with Term Expansion

This chapter reports the empirical evaluation of text categorization with term expansion, using traditional (i.e., without term expansion) text categorization techniques as the performance benchmarks. In the following subsections, the design of the empirical experiments will be detailed, including data collection, evaluation criteria and evaluation procedure. Subsequently, the empirical evaluation results will be discussed.

4.1 Evaluation Design

4.1.1 Data collection
Two document sets were collected for this evaluation, one for term association creation and the other for text categorization. The document set for constructing term associations was collected from a news website (i.e., http://biz.yahoo.com/me/), covering 1,061 business news related to 95 companies from January 2001 to May 2002. Figure 4.1 shows example of term associations constructed from this document set.
The document set for text categorization experiments, containing news documents from November 1, 1999 to December 31, 1999, was collected from a news website (i.e., http://excite.com). Three particular news categories were selected, including business mergers, business partnership, and new product development or release, which together consisted of 408 news documents. A summary of these two document sets is provided in Table 4.1.

Table 4.1. Summary of Data Corpus

<table>
<thead>
<tr>
<th>Document Set</th>
<th>Number of News Documents</th>
<th>Average Number of Words in Each Document</th>
</tr>
</thead>
<tbody>
<tr>
<td>For Term Association Construction</td>
<td>1061</td>
<td>147</td>
</tr>
<tr>
<td>For Text Categorization Experiments</td>
<td>408</td>
<td>135</td>
</tr>
<tr>
<td>Business merger category</td>
<td>244</td>
<td>131</td>
</tr>
<tr>
<td>Business partnership category</td>
<td>86</td>
<td>143</td>
</tr>
<tr>
<td>New product development or release category</td>
<td>78</td>
<td>139</td>
</tr>
</tbody>
</table>
4.1.2 Evaluation Criteria
We measure the effectiveness of a text categorization technique by recall and precision rates. For each category $C$, the recall rate for $C$ is defined as “the percentage of the documents pertaining to $C$ that is correctly assigned to $C$ by the text categorization technique.” On the other hand, the precision rate for category $C$ is defined as “the percentage of documents assigned to $C$ by the text categorization technique that actually belong to $C$.” Given a document set for testing, the aggregate recall (or precision) rate is the average of recall (or precision) rates attained by each category.

4.1.3 Evaluation Procedure
Since a text categorization technique (with or without term expansion) involves a training process, the document set for text categorization experiments needs to be decomposed into two sets: training and testing. To appropriately evaluate the effectiveness of text categorization, the bootstrapping approach [LFM99] was adopted. Specifically, we randomly select 80% of the news documents in the document set for text categorization experiments as a training set and the remaining news documents as a testing set. The training set was then used for learning text categorization patterns by each text categorization technique investigated. Subsequently, the testing set was used for testing the text categorization effectiveness of these techniques. To avoid the bias resulted from random selection and to obtain reliable performance estimates, this train-and-test process was performed thirty times and the overall recall and precision rates of each text categorization technique investigated were estimated by averaging the recall and precision rates obtained from the 30 individual train-and-test processes.
4.2 Evaluation Result

Choice of feature selection method, the number of features \( k \) selected for representing source (input) documents, document representation method, and induction method may affect the effectiveness of text categorization. As mentioned, we implemented correlation coefficient, chi-square (\( \chi^2 \)), and TF×IDF feature selection methods and designed empirical evaluation for performance comparison. In addition, we also examined different numbers of features \( (k) \) to be used for document representation, including 25, 50, 100, 150, 200, 250, and 300. We adopted binary and within document term frequency (TF) schemes for document representation. As for the induction method, we adopted the commonly used symbolic classification analysis techniques—C4.5 and CN2.

4.2.1 Effects of Number of Features on Categorization Effectiveness

We first discuss the effects of number of features selected and used to represent input documents. For any traditional (i.e., without term expansion) text categorization technique, the best categorization result measured by the recall and precision rates generally occurred when the number of features was 25 across different feature selection methods and document representation schemes. Specifically, when C4.5 was employed as the induction method, the resulting precision and recall rates decreased as the number of features increased (as illustrated in Figure 4.2 and Figure 4.3). Similarly, using CN2 as the induction method for traditional text categorization techniques, the precision rate deteriorated as the number of features increased. However, their recall rates varied irregularly the increase of the number of features selected (see Figure 4.4 and Figure 4.5 as examples).
Figure 4.2. Precision Rates of Traditional Text Categorization Techniques (Using Binary Representation Scheme and C4.5 as Induction Method)

Figure 4.3. Recall Rates of Traditional Text Categorization Techniques (Using Binary Representation Scheme and C4.5 as Induction Method)
For text categorization with term expansion, the best categorization result measured by recall and precision rates achieved when the number of features was higher (i.e.,
Regardless whether C4.5 or CN2 was adopted as the induction method for text categorization with term expansion, the resulting precision and recall rates across different feature selection methods and document representation schemes, increased when the number of features increased (see Figure 4.6 and Figure 4.7 as examples, where CN2 was the induction method). This trend was opposite to that exhibited by traditional text categorization techniques.

Figure 4.6. Precision Rates of Text Categorization with Term Expansion (Using TF Representation Scheme and CN2 as Induction Method)
Figure 4.7. Recall Rates of Text Categorization with Term Expansion (Using TF Representation Scheme and CN2 as Induction Method)

4.2.2 Effects of Induction Method, Feature Selection Method and Representation Scheme on Categorization Effectiveness

Adopting the most appropriate number of features for each text categorization technique, the performance achieved by different text categorization techniques is shown in Table 4.2 and Table 4.3, using C4.5 and CN2 as the induction method, respectively. Evidently, the use of C4.5 resulted in higher precision rates but lower recall rates than CN2 in all cases investigated (i.e., regardless whether term expansion was incorporated during text categorization).
Table 4.2. Categorization Effectiveness (Using C4.5 as Induction Method)

<table>
<thead>
<tr>
<th>Feature Selection</th>
<th>Representation Scheme</th>
<th>Without Term Expansion</th>
<th>With Term Expansion</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>$k^*$</td>
<td>Precision</td>
</tr>
<tr>
<td>CC$^1$</td>
<td>Binary</td>
<td>25</td>
<td>59.31%</td>
</tr>
<tr>
<td>CC</td>
<td>TF</td>
<td>25</td>
<td>62.02%</td>
</tr>
<tr>
<td>CS$^2$</td>
<td>Binary</td>
<td>25</td>
<td>67.34%</td>
</tr>
<tr>
<td>CS</td>
<td>TF</td>
<td>25</td>
<td>68.31%</td>
</tr>
<tr>
<td>TF×IDF</td>
<td>Binary</td>
<td>25</td>
<td>61.39%</td>
</tr>
<tr>
<td>TF×IDF</td>
<td>TF</td>
<td>25</td>
<td>63.49%</td>
</tr>
</tbody>
</table>

1: correlation coefficient.
2: Chi square ($\chi^2$).
3: $k^*$ refers to the most appropriate number of features.
4: boldfaced if text categorization with term expansion outperformed its counterpart.

Table 4.3. Categorization Effectiveness (Using CN2 as Induction Method)

<table>
<thead>
<tr>
<th>Feature Selection</th>
<th>Representation Scheme</th>
<th>Without Term Expansion</th>
<th>With Term Expansion</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>$k^*$</td>
<td>Precision</td>
</tr>
<tr>
<td>CC</td>
<td>Binary</td>
<td>25</td>
<td>58.38%</td>
</tr>
<tr>
<td>CC</td>
<td>TF</td>
<td>25</td>
<td>58.25%</td>
</tr>
<tr>
<td>CS</td>
<td>Binary</td>
<td>25</td>
<td>63.53%</td>
</tr>
<tr>
<td>CS</td>
<td>TF</td>
<td>25</td>
<td>63.00%</td>
</tr>
<tr>
<td>TF×IDF</td>
<td>Binary</td>
<td>50</td>
<td>54.76%</td>
</tr>
<tr>
<td>TF×IDF</td>
<td>TF</td>
<td>250</td>
<td>54.70%</td>
</tr>
</tbody>
</table>

The evaluation results of traditional text categorization techniques showed that the feature selection method using the Chi-square ($\chi^2$) metric accomplished the best performance measured by recall and precision rates, across all document presentation schemes and induction methods investigated. The correlation coefficient selection method resulted in the worst performance when C4.5 induction method was employed, while the TF×IDF selection method was least effective in the case of CN2 as induction method. The effects of selection method on categorization effectiveness for text categorization techniques with term expansion are quite different those for traditional text categorization techniques. The correlation coefficient selection method
generally performed better than its counterparts, but the best performance was achieved when the $\chi^2$ was used as the feature selection method with the number of features as 300, the TF representation scheme, and C4.5 as induction method.

As far as document representation scheme is concerned, in the case of C4.5, the TF representation scheme largely outperformed the binary one both in recall and precision rates for all text categorization approaches (traditional and with term expansion). The same conclusion was also observed when CN2 was used in the text categorization techniques supported with term expansion. However, for traditional text categorization techniques, the superiority of one representation scheme to another in either measure (recall or precision rate) was not conclusive.

4.2.3 Comparative Evaluation of Text Categorization Techniques

As shown in Table 4.2 and Table 4.3, when the text categorization techniques with term expansion using the correlation coefficient feature selection method and the C4.5 induction method, higher precision rates and comparable recall rates were achieved, as compared to those attained by their respective counterparts. If CN2 was employed, the text categorization techniques with term expansion significantly outperformed the traditional techniques when the correlation coefficient feature selection method was used.

With respect to document representation scheme, the binary scheme, across different induction and feature selection methods, generally (with an exception when the correlation coefficient feature selection method was employed) was not effective for the text categorization techniques with term expansion, resulting in lower precision and recall rates than those of the traditional techniques. A plausible reason is that the
association strengths of expanded terms are not considered by the binary scheme during document representation. Thus, if terms with low association strengths are selected by a feature selection method for representing documents, these terms will be treated the same as other terms (original or highly associated terms) by the binary representation scheme. Thus, these terms may introduce noises during learning text categorization patterns and, in turn, impair the categorization effectiveness of the text categorization techniques with term expansion. On the other hand, the text categorization techniques with term expansion arrived at compatible or even better precision rates than the non-expansion techniques, if the TF representation scheme was employed. In this particular representation scheme, the incorporation of term expansion into text categorization has shown its effectiveness in improving the accuracy of text categorization measured by precision rates.

The categorization performance of different techniques is further analyzed at the document category level. As mentioned, three news categories were included in the document set for text categorization experiments. Among them, business merger was the largest category, containing 60.06% of the documents. The business partnership category contained 21.04% and the new product development or release category contained 18.90% of the news documents. Using the correlation coefficient feature selection method, the binary representation scheme, and the C4.5 induction method as an example (shown in Figure 4.8 and Figure 4.9), it is evident that traditional text categorization techniques achieved better categorization performance both in precision and recall measures on the largest category (i.e., business merger), while the text categorization techniques with term expansion were more effective on smaller categories (i.e., business partnership and new product development or release). In fact, similar categorization performance for categories with different sizes could also be
observed across different combinations of feature selection methods, document representation schemes and induction methods.

Figure 4.8. Precision Rates of Individual Categories (Using Correlation Coefficient Selection Method, Binary Representation Scheme, and C4.5 as Induction Method)

BM: Business merger category
BP: Business partnership category
NP: New product development or release category
Figure 4.9. Recall Rates of Individual Categories (Using Correlation Coefficient Selection Method, Binary Representation Scheme, and C4.5 as Induction Method)
Chapter 5

Empirical Evaluation for Term Expansion Embedded Document Clustering

This chapter reports the empirical evaluation of document clustering with term expansion, using traditional (i.e., without term expansion) document clustering techniques as the performance benchmarks. This empirical evaluation used the same document sets (one for term association construction and the other for document clustering experiments) as the text categorization evaluation described in Chapter 4. In the following subsections, the design of the empirical experiments and the empirical evaluation results will be detailed.

5.1 Evaluation Design

5.1.1 Evaluation Criteria

As mentioned, the document set collected for text categorization experiments was used for document clustering. We assume that the three categories (i.e., business mergers, business partnership, and new product development or release) are the true categories for the 408 news documents. Accordingly, the effectiveness of a document clustering technique was measured by the purity, diversity, specificity, cluster recall and cluster precision. Purity of a cluster is defined as the maximum number of documents pertaining to the same underlying true category divided by the total number of documents contained in the cluster [ABS99, WHD02]:

\[ \text{Purity}(c) = \frac{n_c}{N_c} \]

where \( n_c \) denotes maximal number of documents in a discovered cluster \( c \) (generated by a document clustering technique) pertaining to the same true category, and \( N_c \)
denotes the number of documents included in the cluster $c$.

The overall purity of a set of discovered clusters can be derived by taking the weighted average of the purity of individual clusters. Hence,

$$Purity = \sum_c Purity(c) \times \frac{N_c}{N}$$

where $N = \sum_c N_c$ denotes the summation of the documents included in each discovered cluster.

From a coverage perspective, a set of discovered clusters should cover all the true categories. The true category coverage denotes the true category from which a discovered cluster contains the maximal number of documents. Let’s assume a discovered cluster $E$ contains 60% of its documents from the true category $A$ and the remaining from the true category $B$. In the case, the true category coverage of $E$ is $A$ because of its dominance in $E$’s current document collection. Thus, diversity is defined as the portion of true categories that are covered by all the discovered clusters under evaluation [ABS99, WHD02].

$$Diversity = \frac{R}{T}$$

where $R$ denotes the number of distinct true categories covered by the discovered clusters under evaluation, and $T$ denotes the number of true categories.

Specificity is defined as the number of true categories covered by the discovered clusters divided by the total number of the discovered clusters [ABS99, WHD02]. Hence,

$$Specificity = \frac{R}{E}$$
where $E$ is the total number of discovered clusters.

The definition of cluster recall and cluster precision are similar to the measure of recall and precision typically used in information retrieval research. Cluster recall and precision are developed based on the concept of association. Inside a cluster (or a true category), an association is a pair of documents belonging to the same cluster [RC99]. Incorrect associations ($IA$) are those that exist in the true categories but do not exist in the discovered clusters. Accordingly, cluster precision is defined as [RC99]:

$$CP = \frac{A_c}{A_u},$$

where $A_c = A_a - IA$ represents total number of correct associations in discovered clusters and $A_a$ represents the total number of associations in the discovered clusters.

On the other hand, cluster recall is defined as [RC99]:

$$CR = \frac{A_c}{A_m},$$

where $A_m$ represents the total number of associations in the true categories.

### 5.1.2 Evaluation Procedure

Similar to the text categorization evaluation, we randomly selected 80% of the news documents in the evaluation data set. This document subset was then used for estimating the clustering effectiveness of a document clustering technique investigated. To obtain reliable performance estimates, this sampling-and-clustering process was performed thirty times and the overall performance estimates for each document clustering technique were estimated by averaging the performance estimates obtained from the 30 individual sampling-and-clustering processes.
5.2 Evaluation Result

5.2.1 Comparative Evaluation of Document Clustering Techniques
Choice of feature selection method, number of features $k$ selected for representing source (input) documents, document representation method, and clustering method may affect the effectiveness of document clustering. As mentioned, we implemented TF and TF×IDF feature selection methods and designed empirical evaluation for performance comparison. In addition, we also examined different numbers of features ($k$) to be used for document representation, including 25, 50, 100, 150, 200, 250, and 300. We adopted binary, within document term frequency (TF), and TF×IDF schemes for document representation. As for the clustering method, we adopted the CLARA clustering analysis technique, known for its efficiency in clustering.

Effects of number of features selected and used to represent input documents are first discussed. For both without and with term expansion document clustering techniques, the TF feature selection method needed a longer number of features to represent the documents to achieve a better clustering result than the TF×IDF feature selection method. For all document clustering techniques (i.e., without and with term expansion), the cluster recall rates increased as the number of features to represent documents increased (see Figure 6.1 as example). As for traditional document clustering techniques, the resulting purity, diversity, and cluster precision deteriorated as the number of features increased (as illustrated in Figure 5.2, Figure 5.3, and Figure 5.4). Specifically, using the TF selection method for document clustering with term expansion, the resulting diversity, specificity, cluster precision, and cluster recall rates increased as the number of features increased.
Figure 5.1. Cluster Recall Rates of Traditional Document Clustering Technique
(Using TF Selection Method)

Figure 5.2. Purity Rates of Traditional Document Clustering Technique
(Using TF×IDF Selection Method)
Adopting the most appropriate number of features for each document clustering technique, the performance achieved by different document clustering techniques is shown in Table 5.1 and Table 5.2, without and with term expansion, respectively.
Table 5.1. Clustering Effectiveness without Term Expansion

<table>
<thead>
<tr>
<th>Feature Selection</th>
<th>Document Representation</th>
<th>Without Term Expansion</th>
<th>(k^*)</th>
<th>CP</th>
<th>CR</th>
<th>Purity</th>
<th>Diversity</th>
<th>Specificity</th>
<th>C#1</th>
</tr>
</thead>
<tbody>
<tr>
<td>TF</td>
<td>Binary</td>
<td></td>
<td>150</td>
<td>43.97%</td>
<td>87.25%</td>
<td>17.91%</td>
<td>43.33%</td>
<td>63.06%</td>
<td>2.2</td>
</tr>
<tr>
<td>TF</td>
<td>TF</td>
<td></td>
<td>100</td>
<td>43.74%</td>
<td>93.84%</td>
<td>30.32%</td>
<td>41.11%</td>
<td>60.00%</td>
<td>2.07</td>
</tr>
<tr>
<td>TF</td>
<td>TF x IDF</td>
<td></td>
<td>150</td>
<td>43.66%</td>
<td>94.40%</td>
<td>20.59%</td>
<td>40.00%</td>
<td>57.22%</td>
<td>2.1</td>
</tr>
<tr>
<td>TF x IDF</td>
<td>Binary</td>
<td></td>
<td>25</td>
<td>45.58%</td>
<td>87.74%</td>
<td>31.35%</td>
<td>54.44%</td>
<td>55.33%</td>
<td>3.07</td>
</tr>
<tr>
<td>TF x IDF</td>
<td>TF</td>
<td></td>
<td>25</td>
<td>43.95%</td>
<td>94.36%</td>
<td>42.77%</td>
<td>44.44%</td>
<td>56.83%</td>
<td>2.5</td>
</tr>
<tr>
<td>TF x IDF</td>
<td>TF x IDF</td>
<td></td>
<td>25</td>
<td>44.00%</td>
<td>94.35%</td>
<td>36.03%</td>
<td>46.67%</td>
<td>60.83%</td>
<td>2.37</td>
</tr>
</tbody>
</table>

1: C# refers to the average number of clusters generated by the target document clustering technique.

Table 5.2. Clustering Effectiveness with Term Expansion

<table>
<thead>
<tr>
<th>Feature Selection</th>
<th>Document Representation</th>
<th>Without Term Expansion</th>
<th>(k^*)</th>
<th>CP</th>
<th>CR</th>
<th>Purity</th>
<th>Diversity</th>
<th>Specificity</th>
<th>C#1</th>
</tr>
</thead>
<tbody>
<tr>
<td>TF</td>
<td>Binary</td>
<td></td>
<td>300</td>
<td>49.41%</td>
<td>76.63%</td>
<td>32.23%</td>
<td>56.67%</td>
<td>85.00%</td>
<td>2</td>
</tr>
<tr>
<td>TF</td>
<td>TF</td>
<td></td>
<td>300</td>
<td>43.50%</td>
<td>93.16%</td>
<td>25.86%</td>
<td>38.89%</td>
<td>56.67%</td>
<td>2.07</td>
</tr>
<tr>
<td>TF</td>
<td>TF x IDF</td>
<td></td>
<td>250</td>
<td>43.66%</td>
<td>90.05%</td>
<td>23.15%</td>
<td>42.22%</td>
<td>60.56%</td>
<td>2.07</td>
</tr>
<tr>
<td>TF x IDF</td>
<td>Binary</td>
<td></td>
<td>250</td>
<td>45.12%</td>
<td>90.92%</td>
<td>16.42%</td>
<td>44.44%</td>
<td>65.00%</td>
<td>2.77</td>
</tr>
<tr>
<td>TF x IDF</td>
<td>TF</td>
<td></td>
<td>25</td>
<td>43.97%</td>
<td>94.27%</td>
<td>39.59%</td>
<td>46.67%</td>
<td>66.89%</td>
<td>2.13</td>
</tr>
<tr>
<td>TF x IDF</td>
<td>TF x IDF</td>
<td></td>
<td>25</td>
<td>43.87%</td>
<td>94.49%</td>
<td>29.79%</td>
<td>43.33%</td>
<td>63.00%</td>
<td>2.1</td>
</tr>
</tbody>
</table>

1: boldfaced if document clustering with term expansion outperformed its counterpart shown in Table 5.1.

As shown in Table 5.1 and Table 5.2, the overall performance of document clustering with term expansion was comparable to that of traditional document clustering techniques. Specifically, when document clustering with term expansion using TF selection method and binary representation method, higher purity, diversity, specificity, and cluster precision rates were achieved. Besides, the specificity achieved by document clustering with term expansion generally was better than the traditional document clustering technique. This evaluation result suggested that the use of term
associations in supporting document clustering could improve the clustering effectiveness when appropriate feature selection method and document representation scheme were adopted.
Chapter 6
Empirical Evaluation for Event Detection with Term Expansion

This chapter reports the empirical evaluation of the event detection with term expansion described in Section 3.4. A traditional feature-based event detection technique was used as the performance benchmarks. In the following, the design of the empirical experiments will be detailed, including data collection, evaluation criteria and evaluation procedure. Subsequently, the empirical evaluation results will be discussed.

6.1 Evaluation Design
6.1.1 Data Collection
This empirical evaluation used the document set that includes business merger, business partnership, and new product development or release news documents, same as that used for the text categorization evaluation described in Chapter 4. Furthermore, the event contained in each news story in the data set for event detection experiments was identified manually. Table 6.1 lists some examples of events in each news categories (i.e., including). Statistics on this data corpus is summarized in Table 6.2. We used the news stories in November 1999 (actually containing 178 news documents) for term association construction and as old news for event detection, while the news stories from December 1999 were used for detection purpose.
### Table 6.1. Some Examples of Event in Each Topics

<table>
<thead>
<tr>
<th>Event Topics</th>
<th>Examples of Events</th>
</tr>
</thead>
<tbody>
<tr>
<td>Business merger</td>
<td>American Home Products mergers with Warner-Lambert</td>
</tr>
<tr>
<td></td>
<td>Red Hat Inc. merges with Cygnus Solutions</td>
</tr>
<tr>
<td>Business partnership</td>
<td>General Motors partners with commerce One Inc.</td>
</tr>
<tr>
<td></td>
<td>Commerce One Inc. partners with CNET Inc.</td>
</tr>
<tr>
<td>New product development or release</td>
<td>Dishnet introduces DSL in India</td>
</tr>
<tr>
<td></td>
<td>Lycos announces Internet’s most powerful service</td>
</tr>
</tbody>
</table>

### Table 6.2. Summary of Data Corpus for Event Detection Evaluation

<table>
<thead>
<tr>
<th>Event Topics</th>
<th>Number of Events</th>
<th>Number of Documents per Event</th>
</tr>
</thead>
<tbody>
<tr>
<td>Business merger category</td>
<td>173</td>
<td>1.41</td>
</tr>
<tr>
<td>Business partnership category</td>
<td>78</td>
<td>1.10</td>
</tr>
<tr>
<td>New product development or release category</td>
<td>69</td>
<td>1.13</td>
</tr>
</tbody>
</table>

### 6.1.2 Evaluation Criteria

We measure the effectiveness of an event detection technique by the miss (false negative) and false alarm (false positive or fallout) rates. The miss rate is defined as the percentage of that an event detection technique fails to detect a new event. On the other hand, the false alarm rate is defined as the percentage of that an event detection technique fails to detect an old event [APL98].

### 6.1.3 Performance Benchmark

The event detection performance of traditional feature-based event detection techniques was used to provide the desired effectiveness benchmarks. Specifically, the single-pass incremental clustering (INCR) for event detection, proposed by Yang [YPC98, YCB99], was employed. We modified the linear time-decaying similarity function by changing the time window from the number of prior news stories to the number of days, as shown in formula 2.12 in Section 2.4.
6.1.4 Evaluation Procedure
Since an event detection technique (with or without term expansion) needed some news stories for old events, the data corpus was separated into two subsets. As mentioned, the news stories in November one were used for the old events and those in December 1999 were used for detection. Similar to the previous evaluations, we randomly selected 80% of the news documents in both data set. The old event data subset was then used as the old events of the event detection techniques and the detection subset was used for the detection process. To obtain reliable performance estimates, this sampling-and-detection process was performed thirty times and the overall performance estimates for each event detection technique were estimated by averaging the performance estimates obtained from the 30 individual sampling-and-detection processes.

6.2 Evaluation Result
6.2.1 Parameter Tuning
Parameter tuning is concerned with, for each event detection technique, selecting appropriate parameter values that result the best event detection performance measured by the evaluation criteria described above. In this study, we used a Detection Error Tradeoff (DET) curve to show how miss and false alarm rates respect to each other at various threshold values. A perfect system would have zero misses and zero false alarms, and would have a “curve” at the origin [APL98]. Thus, a better curve would be generally closer to the origin.

**Parameter Tuning for INCR:**
The number of selected features $k$, time window $w$ and novelty threshold $t_n$ are the
three parameters of the INCR technique. We investigated the number of features $k$ ranging from 50 to infinite (INF) (i.e., $k = 50, 100, 150, 200$, and infinite), the time window $w$ ranging from 7 to 60 ($w = 7, 14, 30$, and 60 days), and the novelty threshold $t_n$ ranging from 0 to 1 at 0.01 increments. Detection Error Tradeoff curves of the traditional INCR technique over different parameter settings are shown in Figure 6.1.

![Figure 6.1. Parameter Tuning Experiments for the INCR Technique](image)

As shown in Figure 6.1, the DET curve closest to the origin was achieved as the time window of 30 days with the infinite number of features. Thus, we selected 30 for the time window and infinite for the number of features.

**Parameter Tuning for INCR with Unrestricted Term Expansion:**
Similar to the parameter tuning of the traditional INCR technique, the number of features \( k \), time window \( w \), and novelty threshold \( t_n \) are the three parameters of the detection phase in the INCR technique with the unrestricted term expansion method. The ranges of numbers of features, time windows, and novelty threshold were the same as those used in the INCR technique. The experiment results are shown in Figure 6.2.

![Figure 6.2. Parameter Tuning Experiments for the INCR Technique with Unrestricted Term Expansion Method](image)

According to Figure 6.2, the DET curve closest to the origin obtained when the time window was 30 days and the number of features was 50. Thus, we selected 30 for the time window and 50 for the number of features.
**Parameter Tuning for INCR with Bounded Term Expansion:**

The number of features $k$, time window $w$, and novelty threshold $tn$ are the three parameters of the detection phase in the INCR technique with the bounded term expansion method. The ranges of numbers of features, time windows, and novelty threshold were the same as those used in the INCR technique. The experiment results are shown in Figure 6.3.

As shown in Figure 6.3, the DET curve was getting closer to the origin as the time window grew from 7 to 60 and the number of features decreased from infinite (INF) to 50. Thus, we selected 60 for the time window and 50 for the number of features.
6.2.2 Comparative Evaluation of Event Detection Techniques

The traditional feature-based event detection (INCR), INCR with unrestricted term expansion and INCR with bounded term expansion techniques were compared using the optimal parameter values tuned in the previous subsection. As can be seen in Figure 6.4, the traditional INCR technique resulted in the best error tradeoff with 10.1% miss rate and 13% false alarm rate (the distance to the origin is 0.1647) at the time window of 60 and novelty threshold of 0.12. The INCR techniques with unrestricted expansion method resulted in the second best error tradeoff with the 17.4% miss rate and the 17.2% false alarm rate (the distance to the origin is 0.2454) at the number of features of 50, time window of 30 days and novelty threshold of 0.12. The INCR technique with bounded expansion method achieved the worst performance. This evaluation result suggested that the usage of term expansion created more noises when supporting event detection.

![DET Curves of Different Event Detection Techniques](image)

Figure 6.4. DET Curves of Different Event Detection Techniques
Chapter 7

Conclusion and Future Research Directions

The rapid accumulation of textual documents on the Web or within an organization requires effective document management techniques, covering from information retrieval, information filtering and text mining. The word mismatch problem represents a challenging issue to be addressed by document management research. Word mismatch has been extensively investigated in information retrieval (IR) research by the use of term expansion (or specifically query expansion in the context of IR). However, a review of text mining literature suggests that the word mismatch problem has seldom been addressed by text mining techniques. Thus, this thesis aims at investigating the use of term expansion on text mining techniques, specifically including text categorization, document clustering and event detection.

In this thesis, we developed term expansion extensions to existing text categorization, document clustering and event detection techniques. The empirical evaluation results showed that term expansion increased the categorization effectiveness when the correlation coefficient feature selection was employed. With respect to document clustering, techniques extended with term expansion achieved comparable clustering effectiveness to existing techniques and showed its superiority in improving clustering specificity measure. Moreover, the use of TF as feature selection method and the binary representation scheme, document clustering with term expansion outperformed traditional document clustering ones. Finally, the use of term expansion for supporting event detection has degraded the detection effectiveness as compared to the traditional event detection technique.
Some future research works related to this study should be continued, including:

1. Quality of term associations might impact the effectiveness of text mining techniques with term expansion. In this study, only the TF×IDF scheme was employed and evaluated. Empirical comparison of other term association construction techniques (e.g., the similarity thesaurus proposed by Qiu and Frei, 1993) is essential.

2. In the text categorization empirical evaluation, only C4.5 and CN2 were adopted as the induction methods. Other induction methods, such as k-nearest neighbor classification and neural network (e.g., backpropagation neural network [RHW86]) should be incorporated in future evaluation.

3. For document clustering, the hierarchical clustering approach (e.g., Hierarchical Agglomerative Clustering [V86]) and the neural-network-based clustering approach (e.g., Kohonen neural network [K89, K95]) were also commonly employed as the clustering methods for document clustering. Empirical evaluation of the effectiveness based on these clustering methods for document clustering with term expansion is desired.

4. The data set used for term association construction consists of some 1,000 news documents. The use of a larger document corpus is needed in future evaluation.

5. The evaluation data set employed in this study comprised news documents across two months and of three categories (or event topics). Larger data sets with more categories (or event topics) and documents for empirical evaluation of the proposed techniques are essential and desired.

6. In this study, we only investigated the use of the term expansion on text categorization, document clustering, and event detection. This research can be extended to investigate the use of term expansion on other text mining techniques (e.g., information extraction).
References


[NH94] Ng, R. and Han, J., “Efficient and Effective Clustering methods for spatial Data Mining,” *Proceedings of International Conference on Very Large Data Bases*, Santiago, Chile, September 1994, pp.144-155.


SIGIR Conference on Research and Development in Information Retrieval, 1996, pp.4-11.


