Knowl Inf Syst (2011) 27:141–161
DOI 10.1007/s10115-010-0296-x

REGULAR PAPER

Extraction, selection and ranking of Field Association (FA) Terms from domain-specific corpora for building a comprehensive FA terms dictionary

Tshering Cigay Dorji · El-sayed Atlam · Susumu Yata · Masao Fuketa · Kazuhiro Morita · Jun-ichi Aoe

Received: 19 May 2009 / Revised: 11 January 2010 / Accepted: 3 April 2010 /
Published online: 24 April 2010

Abstract Field Association (FA) Terms—words or phrases that serve to identify document fields are effective in document classification, similar file retrieval and passage retrieval. But the problem lies in the lack of an effective method to extract and select relevant FA Terms to build a comprehensive dictionary of FA Terms. This paper presents a new method to extract, select and rank FA Terms from domain-specific corpora using part-of-speech (POS) pattern rules, corpora comparison and modified *tf-idf* weighting. Experimental evaluation on 21 fields using 306 MB of domain-specific corpora obtained from English Wikipedia dumps selected up to 2,517 FA Terms (single and compound) per field at precision and recall of 74–97and 65–98. This is better than the traditional methods. The FA Terms dictionary constructed using this method achieved an average accuracy of 97.6% in identifying the fields of 10,077 test documents collected from Wikipedia, Reuters RCV1 corpus and 20 Newsgroup data set.

Keywords Field Association (FA) Terms · Terms weighting and selection · Document classification · Terminology extraction · Information retrieval

T. C. Dorji (✉) · El-sayed Atlam · S. Yata · M. Fuketa · K. Morita · Jun-ichi Aoe
Department of Information Science and Intelligent Systems, Faculty of Engineering,
University of Tokushima, Minamijosanjima 2-1, Tokushima 770-8506, Japan
e-mail: cigay@is.tokushima-u.ac.jp

El-sayed Atlam
e-mail: atlam@is.tokushima-u.ac.jp

S. Yata
e-mail: yata@is.tokushima-u.ac.jp

M. Fuketa
e-mail: fuketa@is.tokushima-u.ac.jp

K. Morita
e-mail: kam@is.tokushima-u.ac.jp

Jun-ichi Aoe
e-mail: aoe@is.tokushima-u.ac.jp

1 Introduction

With the exponential growth of digital texts in recent years, it remains a challenge to retrieve and process this vast amount of unstructured data into useful information and knowledge. As opposed to the traditional methods based on vector space models [21,23] and probabilistic methods [11], novel techniques based on Field Association (FA) Terms [10,26,35] have been found to be very effective in document classification [10], similar file retrieval [1] and passage retrieval [17]. These techniques also hold much promise for application in many other areas such as domain-specific ontology construction [24], machine translation [20], text clustering [13], cross-language retrieval [19], etc.

The concept of Field Association (FA) Terms is based on the fact that the subject of a text (document field) can usually be identified by looking at certain specific words or phrases in that text. It is natural for people to identify the field of a document when they notice these specific words or phrases. These specific words or phrases are called FA Terms. An FA Term is defined as the minimum word or phrase that serves to identify a particular field [10]. FA Terms form a limited set of discriminating terms that can specify document fields [26,32]. For example, "homerun" indicates the subfield <Baseball> of super-field <Sports>, and "US presidential election" indicates sub-field <Election> of super-field <Politics>. Therefore, "homerun" and "US presidential election" are examples of FA Terms.

The FA Terms are attached to a field tree which is a schematic representation of relationships among document fields. The field tree along with the attached FA Terms represents a comprehensive knowledge base. Based on their scope and strength to identify a particular field, FA Terms are classified into levels and given weights.

Presently, one of the main problems lies in the lack of an effective method to build a comprehensive FA Terms dictionary. Although [3] have proposed a method to select compound FA Terms from a pool of single FA Terms, traditional methods [2,3,10,32] have offered no other methods to extract compound FA Terms (FA Terms consisting of more than one word) directly from a document or a corpus. This is a drawback because compound FA Terms form a majority of the relevant FA Terms in a given field. Moreover, the traditional methods for the selection of FA Terms do not use any linguistic information and rely too heavily on the term frequency.

Therefore, in this paper, we present a new methodology that uses both statistical and linguistic methods to extract and select relevant compound as well as single FA Terms from domain-specific corpora. Experimental evaluation shows that the new approach is effective for building a comprehensive FA Terms dictionary as it can extract and select a large number of relevant FA Terms at high precision and recall.

In the rest of the paper, Sect. 2 presents the background and the drawbacks of the traditional methods, Sect. 3 presents the related works, Sect. 4 presents our new methodology and Sect. 5 presents the experimental evaluation. Finally, Sect. 6 presents the conclusion and future work.

2 Background

2.1 FA term

2.1.1 Definition

FA term: FA Terms are words or phrases that allow humans to recognize intuitively the field to which a text belongs. Technically, an FA Term is defined as the minimum word or phrase that can identify a field in a document field representation scheme called field tree.

For example, “First Lady” and “election” are proper FA Terms of the super-field <Politics>, but “current First Lady” and “recent election” are not considered as FA Terms because these terms belong to the same fields as “First Lady” and “election”. The addition of the words “current” and “recent” does not add any new field information. On the other hand, “tournament” is a proper FA Term of super-field <Sports> and “tennis tournament” is a proper FA Term of subfield <Tennis> under super-field <Sports>. In this case, the addition of the word “tennis” has added new field information to the word “tournament”.

Single FA term: A single FA Term is formed by “independent, meaningful, inseparable and smallest unit” that cannot be divided further without losing its semantic meaning [10] usually consisting of a single word. In this paper, two or more words separated by hyphens like “multi-party” but not by white spaces are extracted as a single FA Term.

Compound FA term: A compound FA Term is defined as an FA Term that consists of more than one word. In this paper, only terms consisting of words separated by white spaces are extracted as compound FA Terms. For example, “coalition government” is a compound FA Term of super-field <Politics>.

2.1.2 Comparison with traditional terms

By traditional terms, we mean either index term or terminology. An index term is a term that captures the essence of the topic of a document and is normally used in information retrieval [14]. Index terms make up a controlled vocabulary for use in bibliographic records and are used as keywords to retrieve documents in an information system like a catalog or a search engine.

On the other hand, terminology denotes a more formal discipline which systematically studies the labeling or designating of concepts particular to one or more subject fields or domains of human activity, through research and analysis of terms in context, for the purpose of documenting and promoting correct usage [41]. It is commonly used in translation and for representing knowledge in a particular domain. However, sometimes the terms ‘terminology’ and ‘index term’ are used interchangeably as in Saneifar et al. [29].

FA Terms offer a more formalized study of terms that can discriminate between different fields. FA Terms belong to fields in a classification scheme called field tree, and the purpose of FA Terms is to create a comprehensive knowledge base. At the level of individual terms, FA Terms have similarities to both index terms and terminologies, but they are also different in some ways. FA Terms are similar to index terms in that they both consist of choosing a set of statistically and contextually significant terms including names of people and places etc. used in a particular document. But index terms are less strictly defined than FA Terms, and the choice of index terms may depend a lot on the user and the purpose for which they are selected. Unlike FA Terms, sometimes index terms may also consist of word stems.

The formal study of terminology may exclude names of people and places that qualify as FA Terms. Similarly, some subject terminologies which do not occur in documents would not qualify as FA Terms. Also, the traditional terms are usually managed in isolation for each domain [12] unlike the FA Terms which are managed as a knowledge base.

FA Terms are extracted from a document by taking their occurrence both in the document and in the whole domain specific corpus into consideration, while index terms may be extracted directly from individual documents. Thus FA Terms have high field specificity while the same cannot be said of index terms. In addition, as pointed out in Sect. 2.1.1, by definition, FA Terms are restricted to the minimum word or phrase that can identify a document field. So, while “election” is a FA Term, “recent election” is not one. There are no such restrictions on the index term and terminology.

Fig. 1 Snapshot of field tree structure

```
01. Culture & Fine Arts
  01.0 GENERAL
   --------
     ------
  01.5 ARCHITECTURE
  01.6 MUSIC
     01.6.0 General
     01.6.1 Classic
     01.6.2 Popular
06. Hobby & Entertainment
  ----
  06.5 MOTOR VEHICLES
  06.6 HORSE RACING
```

2.2 Field tree

A document field is defined as basic and common knowledge useful for human communication [10]. A field tree is a schematic representation of relationships among document fields. Leaf nodes in the field tree correspond to terminal fields, nodes connected to the root are super-fields and other nodes correspond to median fields. FA Terms are saved in a field tree as knowledge base. This knowledge base would constitute an FA Terms dictionary.

In this study, we use a field tree based on Imidas'99 [8] containing 14 super-fields, 50 median fields and 393 terminal fields (sub-fields). For example, in Fig. 1, the path <Culture & Fine Arts/Music/Classic> describes super-field <Culture & Fine Arts> having median-field <Music> and terminal field <Classic> and this path can be represented by field code 01.6.1.

Each FA Term is connected to a particular field inside a hierarchical field tree like the one shown in Fig. 1. Since an FA Term may belong to more than one field, it is possible that the same FA Term may be connected to the field tree at more than one node. In the FA Terms dictionary, whether an FA Term belongs to more than one field or not is represented by its level.

2.3 FA term levels

Some FA Terms can uniquely identify a certain field, while some FA terms may belong to two or more fields. Thus, each FA Term has a different scope to associate with a field. In order to take this into consideration, FA Terms are classified into five different levels [10] based on how well they indicate specific fields as shown in Table 1. The FA Term levels are defined as follows:

Level 1: Proper FA Terms-terms associated with one subfield only.
Level 2: Semi-proper FA Terms—terms associated with more than one subfield in one super-field.
Level 3: Super FA Terms—terms associated with one super-field,
Level 4: Cross FA Terms—terms associated with more than one subfield of more than one super-field
Level 5: Non FA Terms—terms that do no specify any subfield or super-field.

Table 1 Examples of FA Terms with their levels

FA Term	Field	Level
Four-wheel drive	Sub-field <Motor vehicles>	1
Tournament	Sub-fields <Soccer>, <tennis>, <basketball> etc. under super-field <Sports>	2
Political party	Super-field <Politics>	3
Victory	Sub-field <election> under super-field <Politics> Sub-field <Tennis> under super-field <Sports>	4
Huge size	No particular field	5

2.4 Drawbacks of the existing methods

Fuketa et al. [10] relied on the extraction of FA Terms from a manually collected document corpus using “weighted inverse document frequency (WIDF)” which was defined as the term frequency of a word in a sub-field divided by the term frequency of the word in the whole corpus. Although this method was found useful in selecting FA Terms on a small scale in Japanese, it is not effective for extracting and selecting English FA Terms on a larger scale. Atlam et al. [2] presented a method to extract single FA Terms from the Internet using the search engine. In this method, FA Terms are selected based on the “Concentration Ratio” of the FA Term candidates. The “Concentration Ratio” is calculated using the term frequency of an FA Term candidate in the documents obtained from the Internet using a commercial search engines such as Yahoo or Google. Sharif et al. [32] proposed a method to improve Atlam’s method [2] by using a passage retrieval technique. The only difference between Atlam’s method and Sharif’s method [32] is that Sharif’s method calculates the term frequency of FA Term candidates based on passages retrieved by Salton’s passage retrieval technique [28] rather than using whole document texts. Both these methods [2,32] have the following drawbacks:

- Sorting and compiling a large collection of documents for different fields using commercial search engines would be tedious and mistake-prone.
- Calculation of “Concentration Ratio” for selecting relevant FA Terms uses only term frequency. No other component for domain relevance is used in selecting FA Terms from a pool of FA Term candidates. Term frequency may not be the only determining factor for a term’s association with a field.
- Although [3] have proposed a method to select compound FA Terms from a pool of single FA Terms, this method is constrained by the limitation of the single FA Term extraction methods [2,10,32], since it is not able to extract compound FA Terms directly from the document or the corpus. Since compound FA Terms form a majority of the FA Terms in a given field, automatic extraction of compound FA Terms from a document or a corpus is very important.

In view of the aforementioned drawbacks, these methods [2,10,32] are not suitable for extracting and selecting FA Terms to build a comprehensive FA Terms dictionary. In order to overcome these drawbacks, this paper presents a new methodology using linguistic and statistical techniques to extract and select both single and compound FA Terms automatically from domain-specific corpora obtained from Wikipedia dumps [40].

3 Related works

There are two important aspects to the construction of FA Terms dictionary: the extraction and selection process of FA Terms, and the organization of FA Terms within the dictionary as a knowledge base. These two aspects bear similarities to different areas of research in the field of information science. The former has similarities with the techniques used in terminology extraction, while the latter has similarities with knowledge organization.

Terminology management is a key component of many natural language processing activities such as machine translation (Langlais et al., 2004), text summarization and text indexation [22]. The identification of terminology in the biomedical literature is seen as a very important and challenging research area [15]. With the ever increasing popularity of the Internet, terminology management assumes continued significance for developing content-based applications. For instance, terminology extraction is used to model communities and networked enterprises on the web [37]. The extracted terminologies are also used for applications like topic-driven web crawlers [34], web services and recommender systems [25], conceptualizing knowledge domain, supporting the creation of a domain ontology, etc.

Krauthammer et al. [15] list extensive lexical variations, term synonymy (when a concept is represented with several terms) and term homonymy which creates uncertainties regarding the exact term identity as barriers to successful term identification. The maintenance of terminological resources is further complicated by a constantly changing terminology. Some terms appear in a very short time period, and some of them do not last for long.

The extraction of domain terminology from textual data is an important step for creating a specialized dictionary of terminologies [22,37]. Krauthammer et al. [15] provide an overview of a number of approaches used for term identification. Co-occuring words are usually extracted as compound term candidates. Approaches used to identify co-occurrences consist of dictionary-based, syntactic rule-based using POS Patterns, machine-learning and *n*-gram based approaches, or a hybrid of these approaches. For instance [4,38] have developed syntactic pattern rules for extracting noun phrases, while machine learning methods based on Hidden Markov Models (HMMs) are used in Collier et al. [7] to extract terminology in the field of molecular biology. Saneifar et al. [29] use syntactic patterns as well as bigrams to extract terminology candidates from log files.

Smadja [33] has presented a terminology extraction system called Xtract which first uses straight statistical measures to retrieve from a corpus pairwise lexical relations whose common appearance within a single sentence are correlated. A pair (or bigram) is retrieved if its frequency of occurrence is above a certain threshold and if the words are used in relatively rigid ways. Xtract then uses the output bigrams to extract collocations involving more than two words (or *n*-grams) based on syntactic and statistical information. Sclano and Velardi [31] has presented TermExtractor which first identifies term candidates based on syntactic patterns and then selects the relevant terms using the so-called domain pertinence, domain consensus and lexical cohesion filters.

Statistical methods are generally used with syntactic methods for evaluating the adequacy of terminological candidates. Under the framework established by traditional terminology extraction methods, we use specially developed POS patterns to extract FA Term candidates from domain-specific corpora using a sliding window of ten words. Relevant FA Terms are then selected by corpora comparison and using a unique series of statistical formulae based on *tf-idf*.

The selected FA terms are then added to FA Terms dictionary under their relevant fields in the field tree. Currently, we use a hierarchical field tree based on Imidas'99 [8]. This field tree is a classification system similar to those used in knowledge organization systems. Therefore,

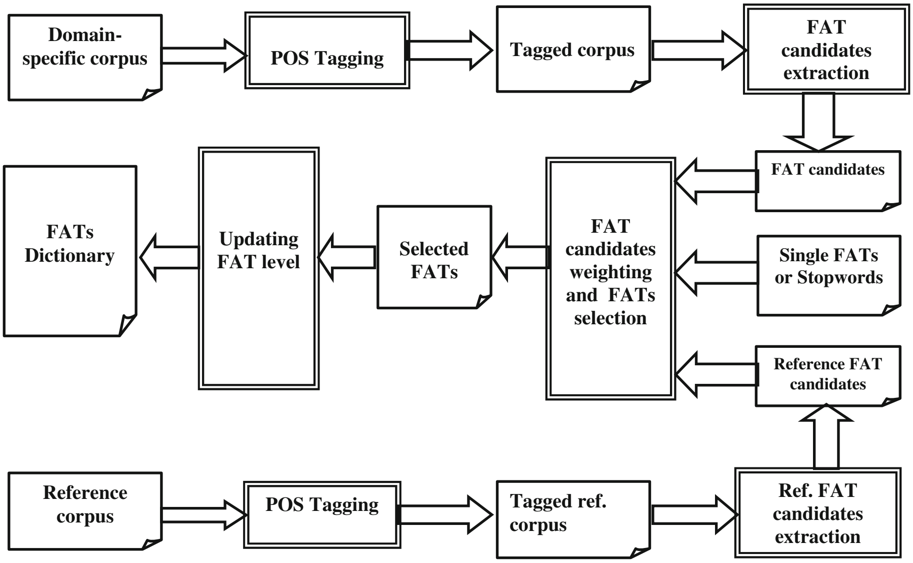

Fig. 2 System outline of the proposed FA Terms selection methodology

if required, it can be updated to other knowledge organization systems such as those based on traditional hierarchical classification systems or faceted analytic approaches [5].

4 The new methodology

4.1 System outline

The outline of the new system for the selection of FA Terms and building a comprehensive FA Terms dictionary is shown in Fig. 2. The abbreviation FAT stands for FA Term. We use as inputs domain-specific corpora for the various fields of interest and reference corpora for comparison. The system consists of a part-of-speech (POS) tagger, a module for FA Terms candidate extraction, a module for weighting candidate terms and selecting the relevant FA Terms, and lastly a module for determining the level of selected FA Terms and appending them to the FA Terms dictionary.

Firstly, documents in a domain-specific corpus are POS tagged using TreeTagger [30,36]. The tagged corpus is then fed as input to the FA Term candidate extractor module. This module extracts FA Term candidates that match predefined POS pattern rules. The extracted FA term candidates are then weighted and ranked by comparing with term candidates from a reference corpus and using formulae based on *tf-idf*. Candidate terms that have normalized final weights higher than the heuristically determined cutoff weight are automatically selected as new FA Terms. The selected FA Terms are then manually checked to confirm their relevance. Finally, the selected FA Terms are compared with all other FA Terms in the dictionary and their FA Term levels are determined by the module for updating FA Term levels. Then the selected FA Terms are appended to the FA Terms dictionary under their relevant fields.

The extraction and selection of single FA Terms and compound FA Terms are done separately because their term frequencies differ drastically. Moreover, stopwords list to filter

Table 2 Sample output of TreeTagger

Token	POS	Lemma	Token	POS	Lemma	Token	POS	Lemma
Red	NP	Red	Van	NP	Van	troops	NNS	troop
,	,	,	Riper	NP	Riper	,	,	,
Commanded	VVN	Command	,	,	,	Evading	VVG	Evade
By	IN	by	Used	VVN	Use	Blue	NP	Blue
Retired	VVN	Retire	Motorcycle	NN	Motorcycle	‘s	POS	‘s
Marine	NP	Marine	Messengers	NNS	Messenger	Sophisticated	JJ	Sophisticated
Corps	NP	Corps	to	TO	to	Electronic	JJ	Electronic
Lt.	NP	Lt.	Transmit	VV	Transmit	Surveillance	NN	Surveillance
General	NP	General	Orders	NNS	Order	Network	NN	Network
Paul	NP	Paul	to	TO	to	.	SENT	.
K.	NP	K.	Front-line	NN	Front-line			

out term candidates is used only during the selection of single FA Terms, whereas single FA Terms list to give additional weights is used only during the selection of compound FA Terms. All these procedures involved in the extraction and selection of FA Terms are described in detail in Sects. 4.2, 4.3, 4.4 and 4.5.

4.2 POS tagging

The documents in the domain-specific corpora and the reference corpora are POS-tagged using TreeTagger [30,36]. Schmid [30] presented a probabilistic tagging method in which transition probabilities are estimated using a decision tree and thus avoids the problems that Markov models based taggers face when they have to estimate transition probabilities from sparse data. TreeTagger is reported to achieve accuracy of 96.36% on Penn-Treebank data which is better than that of a trigram tagger on the same data. It annotates text with POS and lemma information as shown in Example 1.

Example 1 The results of tagging the following sentence using the TreeTagger is shown in Table 2. The sentence was taken from a document in the domain-specific corpus of the ‘military’ field.

> Red, commanded by retired Marine Corps Lt. General Paul K. Van Riper, used motorcycle messengers to transmit orders to front-line troops, evading Blue’s sophisticated electronic surveillance network.

4.3 FA term candidates extraction

4.3.1 Single FA term candidates extraction

Single words like “ombudsman” and two or more words joined by hyphens like “self-determination”, “commander-in-chief” etc. which are common nouns, proper nouns, adjectives or gerunds are extracted as candidates for single FA Terms. The words that belong to these parts-of-speech are the most likely candidates for single FA Terms. We extract the actual word used as well as its lemma. However, we do not replace the actual words used with their lemma because it sometimes leads to the loss of semantic information.

Example 2 Let us consider extracting single FA Term candidates from the sentence given in Example 1. All words, the POS of which have been identified as nouns, adjectives or gerunds in Table 2 would be extracted as candidates. That would include the following words: *Red, Marine, Corps, Lt., General, Paul, K., Van, Riper, motorcycle, messengers, orders, front-line, troops, evading, Blue, sophisticated, electronic, surveillance, network.*

4.3.2 Compound FA term candidates extraction

4.3.2.1 Extraction by matching POS patterns Compound FA Terms are formed by collocations. Smadja [33] identified three types of collocations: rigid noun phrases, predicative relations and phrasal templates. Compound FA Terms consist of an uninterrupted sequence of words such as “parliamentary election”, “Barack Obama”, “Council of Ministers”, “Christian Heritage Party”, “Declaration of the Rights”, etc. and fall under the category of “rigid noun phrases”.

Bennet et al. and Voutilamen [4,38] have developed syntactic rules for extracting noun phrases in general, but they cannot be applied directly for our purpose as we are interested only in some special noun phrases that are candidates for compound FA Terms. All noun phrases cannot be candidates for FA Terms. Based on previous studies [4,33,38] and on our own study of FA Terms, we developed the following sequence of POS patterns for a maximum length of ten words and minimum of two words, as rules for determining compound FA Term candidates:

1. [Noun]–[Noun]—[up to 8 more nouns]
2. [Noun]–[Preposition]–[Noun]—[up to 7 more nouns]
3. [Noun]–[Preposition]–[Article]–[Noun]—[up to 6 more nouns]
4. [Adjective]–[Noun/Gerund]—[up to 8 more nouns]
5. [Adjective]–[Adjective]–[Noun]—[up to 7 more nouns]
6. [Gerund]–[Noun]—[up to 8 more nouns]

These rules are applied to the tagged documents from the corpora using a sliding window of ten words. The window is placed on the words such that the word at the beginning of the window is a noun, adjective or a gerund as per the POS pattern for a compound FA Term candidate identified earlier. The POS pattern rule is then applied to the window contents. The window will be truncated when a word that does not conform to the identified POS pattern is encountered or a punctuation mark other than the hyphen is encountered. Whether a candidate term is located or not, the window then slides over to the next word that matches the starting POS for a FA Term candidate following the word where the previous window was truncated. If the previous window was not truncated, the window moves to the word that matches the starting POS for a FA Term candidate next to where the previous window ended. The process is repeated until the end of the file is reached.

Example 3 Let us look at the extraction of compound FA Term candidates from the sentence given in Example 1. Figure 3 shows the first three steps involved in extracting compound FA Term candidates. The box formed by broken lines show the position of the ten word sliding window at each step, while the box formed by the dark unbroken lines show the position where the window is truncated after identifying a possible FA Term candidate. Firstly, the underlined phrases are extracted as compound FA Term candidates.

> *Red, commanded by retired* ***Marine Corps Lt. General Paul K. Van Riper*** *, used* ***motorcycle messengers*** *to transmit orders to* ***front-line troops*** *,****evading Blue*** *’s* ***sophisticated electronic surveillance network****.*

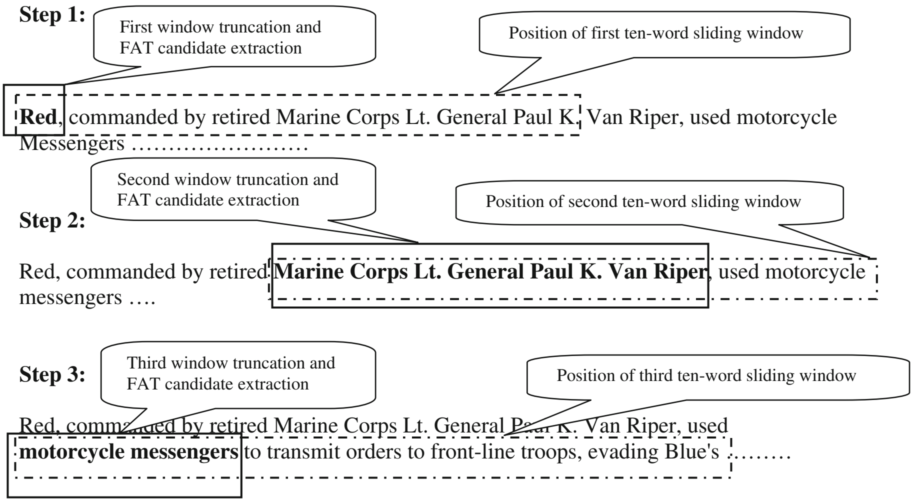
Step 1:
First window truncation and FAT candidate extraction
Position of first ten-word sliding window
Red, commanded by retired Marine Corps Lt. General Paul K. Van Riper, used motorcycle Messengers ……………………
Step 2:
Second window truncation and FAT candidate extraction
Position of second ten-word sliding window
Red, commanded by retired Marine Corps Lt. General Paul K. Van Riper, used motorcycle messengers ….
Step 3:
Third window truncation and FAT candidate extraction
Position of third ten-word sliding window
Red, commanded by retired Marine Corps Lt. General Paul K. Van Riper, used motorcycle messengers to transmit orders to front-line troops, evading Blue's ………

Fig. 3 An example of compound FA Term candidate extraction

Secondly, some of the FA Term candidates can furnish other smaller FA Term candidates, since some of the POS pattern rules are subsets of other rules. Such terms are extracted as described in Sect. *4.3.2.2*. Finally, the following compound FA Term candidates are extracted: *Marine Corps Lt. General Paul K. Van Riper, motorcycle messengers, front-line troops, evading Blue, sophisticated electronic surveillance network, surveillance network, sophisticated electronic surveillance.*

4.3.2.2 Extracting subset FA Term candidates Candidate terms made up of three or more words have the potential to yield smaller FA Term candidates, since some of the POS pattern rules are subsets of other rules. In Example 3, the FA Term candidate "*sophisticated electronic surveillance network"* was extracted at first. Then this FA Term candidate also yielded two smaller FA Terms "*surveillance network"* and"*sophisticated electronic surveillance"*.

Similarly, an FA Term candidate like "*proportional representation system*" which matches the POS pattern "adjective–noun–noun", would furnish two smaller terms "*proportional representation*" and "*representation system*". "*European Parliament elections*" which matches the POS pattern "noun–noun–noun" would furnish two smaller terms "*Parliament elections*" and "*European Parliament*".

On the other hand, the term "*members of the European Parliament*" which matches the POS pattern "noun–preposition–article–noun–noun" would furnish only one smaller term "*European Parliament*" because we do not qualify phrases starting or ending with preposition or article as FA Term candidates. We also do not break down a sequence of proper nouns as they usually refer to names of people or places. Hence the term "*Marine Corps Lt. General Paul K. Van Riper*" was not broken down in our example in *Example 1* in Sect. *4.3.2.1.*

4.4 FA terms weighting and selection

In this section, we describe how the extracted FA Term candidates are weighted and the relevant FA Terms are selected using corpora comparison and statistical formulae based on *tf-idf*. Single FA Term candidates and compound FA Term candidates are weighted separately, since they have drastically different term frequencies. However, the same formulae and procedure are used for both.

4.4.1 Corpora comparison

Comparing the occurrence of a FA Term candidate in the domain-specific corpus with its occurrence in the reference corpus [9,12] is an effective method to find FA Terms with high field specificity. The reference corpus is chosen in such a way that it would help us discriminate FA Terms in the domain-specific corpus more distinctly.

For each candidate term, we measure the local term frequency (the frequency of term within the given document), the global term frequency (term frequency within the whole of domain-specific corpus) and the document frequency (number of documents in the domain-specific corpus that contain the term). Likewise, we also measure the global term frequency and the document frequency of the term in the reference corpus. These measures are then used in the term weighting formula described in Sect. 4.4.3. The more relevant a term is to the field, the higher will be its term frequency and document frequency in the domain-specific corpus, while its term frequency and the document frequency in the reference corpus would be lower. This would help us differentiate genuine FA Term candidates from those terms of general expression which may occur frequently in both domain-specific corpus and reference corpus.

4.4.2 Use of stopwords list and single FA terms list

We use a list of 534 stopwords to filter out the single FA Term candidates during the selection process. The list is not used during the selection of compound FA Terms. The stopwords consist of function words, names of months, days and directions, since they do not identify any field strongly. The following are few examples of the stopwords used: *a, able, about, above, across, after, again, against, all, almost, alone, along, already, also, although, always, among, an, and, another* etc.

On the other hand, we use a list of single FA Terms during the compound FA Term selection process. Since compound FA Terms often contain single FA Terms [3], we refer to the list of single FA Terms during compound FA Terms selection. We give additional weights to those candidate terms containing a single FA Term as shown in Sect. 4.4.3.

4.4.3 Weighting formula

The weighting formula is based on a modified version of *tf-idf.* Lan et al. [16] has made a comparative study of various versions of *tf-idf* term weighting schemes. Brunzel et al. [6] introduced a term weighting scheme which improves the behavior compared to the traditional *tf-idf* scheme by adding a domain relevance component that measures the degree to which a term is regarded as more relevant within a corpus compared to a reference corpus. Based on the careful study of the different term weighting schemes ([6,16,18], we have come up with a unique term weighting method of our own based on a combination of term frequency (*tf*), document frequency (*df*), inverse document frequency (*idf*) and inverse term frequency (*itf*) of a term in a domain-specific corpus and a reference corpus. The use of *itf* was first introduced by Leopold and Kindermann [18] although the calculation of *itf* in this paper is not exactly the same as theirs. The proposed term weighting method was found to be better suited for the selection of FA Terms from domain-specific corpora compared to the traditional methods.

We make the calculations at two levels: one at the document level and the other at the corpus level. “Local” refers to the calculation at the level of the document, while “global” refers to the calculation at the level of the whole corpus of a particular field. The final selection of FA Terms is based on the global weight.

weighted_local_tf refers to weighted local term frequency of a FA Term candidate. *local_tf* refers to local term frequency and *local_avg_tf* to average frequency of all FA term candidates in a document.

$$weighted\_local\_tf = \frac{1 + k1 \times \log(local\_tf)}{1 + k2 \times \log(local\_avg\_tf)} \tag{1}$$

k1 and *k2* are normalizing terms calculated using document frequencies to curb the influence of unusually high term frequency in a particular document. *N1* is the number of documents in the domain-specific corpus, *df1* is the number of documents containing the term in the domain-specific corpus, *N2* is the number of documents in the reference corpus, *df2* is the number of documents containing the term in the reference corpus and β is an adjustment factor.

$$k1 = \frac{df1/N1}{df1/N1 + df2/N2} \times \beta \tag{2}$$

$$k2 = \frac{df2/N2}{df1/N1 + df2/N2} \times \beta \tag{3}$$

Then, *local_term_weight* is calculated as in Eq. (4). α is the additional weight given to compound FA Term candidates if they contain a single FA Term. itf_1, itf_2, idf_1 and idf_2 are calculated as $\log_{10}(Nt_1/tf_1)$, $\log_{10}(Nt_2/tf_2)$, $\log_{10}(N_1/df_1)$ and $\log_{10}(N_2/df_2)$ respectively. Nt_1 and tf_1 are the total number of candidate terms and the total frequency of a particular term in the domain-specific corpus, while Nt_2 and tf_2 are the total number of candidate terms and the total frequency of a particular term in the reference corpus.

$$local\_term\_weight = weighted\_local\_tf \times \left(\frac{itf_2}{itf_1} \times \frac{idf_2}{idf_1}\right)^2 + \alpha \tag{4}$$

After calculating *local_term_weight*, candidate terms which do not make it beyond a heuristic cut-off weight are filtered out. Global term weights are calculated as in Eq. (5) by taking the average of the local term weights that remain.

$$global\_term\_weight = \frac{\sum_{i=1}^{n} local\_term\_weight_i}{n} \tag{5}$$

Where *n* is the number of documents in which the term appears in the domain-specific corpus.

The *global_term_weight* calculated earlier is often spread over a wide range. In order to get normalized weights between 0 and 1, the FA Term candidates are then ranked by giving the value of 1 for the term with $global\_term\_weight_{\max}$ the highest *global_term_weight,* and so on. For the *i*th term, normalized final weight is calculated as shown in Eq. (8). *lowest_rank* refers to the lowest rank value.

$$normalized\_rank_i = \frac{lowest\_rank + 1 - rank_i}{lowest\_rank + 1} \tag{6}$$

$$normalized\_wt_i = \frac{global\_term\_weight_i}{global\_term\_weight_{\max}} \tag{7}$$

Table 3 Example showing FA Term weight calculation

FA term candidate	*df1*	*df2*	*tf1*	*tf2*	*N*1	*N*2	*Nt*1	*Nt*2	*global_term_weight*	*Rank*	*Normalized_final_weight*
Volkswagen polo	3	0	7	0	244	300	9,520	9,632	167.507	163	0.6774
Sports car	36	1	63	3	244	300	9,520	9,632	318.82	22	0.8622

$$normalized\_final\_weight_i = \frac{\lambda \times normalized\_wt_i + \mu \times normalized\_rank_i}{\lambda + \mu} \tag{8}$$

λ and μ are adjustment factors. The FA Term candidates with *normalized_final_weight* above a heuristically determined cut-off value are automatically selected as FA Terms.

Example 4 Let us take the FA Term candidate ‘Volkswagen Polo’ for the field <Motor vehicles> for instance. As Table 3 shows, it appears 7 times in 3 documents of the domain specific corpus, but it does not appear at all in the reference corpus.

First we need to calculate *weighted_local_tf* by Eq. (1). Suppose in the first document, ‘Volkswagen Polo’ appears 2 times (*local_tf* = 2), and the average frequency of all terms in the document is 1.2 (*local_avg_tf* = 1.2). From Table 1, *df1* = 3, *N1* = 244, *N2* = 300, and *df2* = 0.‘001 (Set to 0.001 to avoid division by 0). Given $\beta = 10$, $k1 = 9.997$ and $k2 = 0.003$. Thus, from Eq. (1), *weighted_local_tf* is found to be 3.999.

Using the parameter values (using 0.001 instead of 0) from Table 1, we get *itf1* = 3.134, *itf2* = 6.984, *idf1* = 1.910, *idf2* = 5.477. Now, from Eq. (4), we get *local_term_weight* = 167.915, where we have used $\alpha = 5$, since it contains a valid single FA Term ‘Volkswagen’. Likewise, we calculate *local_term_weight* for all the 3 documents in which ‘Volkswagen Polo’ appears.

Finally, *global_term_weight* is calculated by taking the average of the 3 local term weights as shown in Eq. (5), and is found to be 167.501. It is ranked 163 out of 2,515 FA Term candidates. So, in Eq. (6), $lowest\_rank$ = 2,515 and $rank_i = 163$. In Eq. (7), $global\_term\_weight_i = 167.501.global\_term\_weight_{\max}$ was found to be 575.00. Given $\lambda = 2$ and $\mu = 3$, we get *normalized_final_weight* for ‘Volkswagen Polo’ to be 0.6774 from Eq. (8). Likewise, the *normalized_final_weight* of the FA Term ‘Sports car’ is calculated to be 0.8662.

4.5 Updating FA terms level

Before appending the selected FA Terms to the FA Terms dictionary, their levels are determined by comparing them with the existing FA Terms in the dictionary. If the same FA Term already exists under the same field, it is not appended. If the term does not already exist in the dictionary, it is temporarily designated as a proper FA Term. If the same term exists under a different subfield under the same super-field, the new FA Term is designated as level 2, semi-proper FA Term. If the term belongs to a super-field and it does not already exist in the dictionary, it is designated as level 3, super FA Term. If the same term exists in the subfield of another super-field, the new FA Term is designated as level 4, cross FA Term. Using this method, the accuracy of FA Term levels would increase as the number of FA Terms increases.

5 Experimental evaluation

The experimental procedure follows the system outline presented in Fig. 1. The inputs are domain-specific corpora, and reference corpora for comparison. We discuss these corpora in Sect. 5.1. We evaluate the results of the extraction and selection of FA Term candidates from the domain-specific corpora in Sect. 5.2. The effectiveness of the presented methodology is measured by precision and recall of FA Terms selection. Finally, we discuss the building of FA Terms Dictionary from the selected FA Terms, and evaluating its accuracy in Sect. 5.3. The accuracy of the FA Terms dictionary is tested by its ability to identify the fields of test documents collected from Wikipedia dumps, Reuters RCV1 corpus [27] and 20 Newsgroup data set (downloaded from http://people.csail.mit.edu/jrennie/20Newsgroups/on1October2009). The experimental results are discussed in Sect. 5.4.

5.1 Corpora collection

The domain-specific corpora used in this research were collected from the English Wikipedia dump [40] downloaded on 24 July 2008. As [39] have shown that a thesaurus of concepts built from Wikipedia is effective in enhancing previous approaches for text classification, Wikipedia dumps is a good source of corpora for extracting FA Terms. From the downloaded Wikipedia dumps, the individual documents (articles) are extracted and POS-tagged with TreeTagger. We then use these tagged corpora as the source of our domain-specific corpora. We divide the documents (articles) into different fields based on their Wikipedia categories and titles using a computer program. Some manual checking was required to get rid of garbage or empty files. The size of each domain-specific corpus is shown in Table 4.

Experimental evaluation is carried out for 21 different fields using 306.28 MB of domain-specific corpora. The 21 fields are: <Motor vehicles>, <Music>, <Films>, <Politics>, <Military>, <Christianity>, <Aviation>, <Sex>, <Geography>, <Telecommunications>, <Soccer>, <History>, <Economics>, <Transport>, <Education>, <Foreign relations>, <Computer software>, <War>, <Mathematics>, <Sports>, and <Television>. Reference corpora for comparison were also collected from the Wikipedia dump.

5.2 Evaluation of FA terms selection results

Once the domain-specific corpora and reference corpora are ready, we extract the single FA Terms and compound FA Terms as described in Sect. 4.3. The number of candidates selected in different fields is shown in Table 4. The extracted FA Term candidates are then given weights using the method described in Sect. 4.4.3. The FA Term candidates with *normalized_final_weight* above a heuristic threshold value are selected as FA Terms. Based on our experimental observations, we selected $\alpha = 5$, $\beta = 10$, $\lambda = 2$, and $\mu = 3$. Out of the total number of FA Term candidates, only a very small percentage is selected as FA Terms.

The selection of FA Terms is done based on the *normalized_final_weight* which lies between 0 and 1. The column 'Cutoff weight' in Table 4 shows the *normalized_final_weight* threshold above which the FA Term candidates were selected as FA Terms for different fields as shown in Table 4.

For space reasons, in Table 4, we show the results of FA Terms selection for only 10 fields out of the 21 used in the experimental evaluation. The column 'Total FA Terms' shows the total number of FA Terms (single and compound separately) selected for the field after relevant terms missed by the program were manually added. The column "FA Terms selected automatically" refers to the number of FA Terms selected automatically by the system. The

Table 4 Results of FA Terms selection for some fields

Field (Size in MB)	FAT* type	Candidate terms	Total FA terms	FA Terms selected automatically	Cutoff weight	Irrelevant terms selected	Precision %	Recall %
Motor vehicles (6.32)	SFAT*	177,047	1,113	1,077	0.43	85	92.69	96.77
	CFAT*	79,155	1,404	1,266	0.42	60	95.48	90.17
Military (18.20)	SFAT	556,566	338	228	0.59	32	87.69	67.46
	CFAT	221,230	1,327	868	0.49	103	89.39	65.41
Christianity (13.30)	SFAT	367,558	377	286	0.50	57	83.38	75.86
	CFAT	129,305	1,211	1,152	0.44	107	91.50	95.13
Music (18.30)	SFAT	498,529	622	514	0.54	21	96.07	82.64
	CFAT	201,899	1,345	1,247	0.42	145	89.58	92.71
Politics (31.70)	SFAT	839,412	294	236	0.57	69	77.38	80.27
	CFAT	323,930	1,858	1,716	0.36	84	95.33	92.36
Computer software (8.14)	SFAT	220,163	480	442	0.54	34	92.86	92.08
	CFAT	95,798	1,382	1,285	0.42	46	96.54	92.98
History (29.30)	SFAT	784,655	184	160	0.61	52	75.47	86.96
	CFAT	287,847	313	243	0.66	47	83.79	77.64
Telecommunications (6.60)	SFAT	162,876	398	320	0.54	87	78.62	80.40
	CFAT	72,661	1086	1,051	0.43	82	92.60	93.72
Soccer (2.89)	SFAT	71,477	371	309	0.55	60	83.74	83.29
	CFAT	33,295	908	851	0.47	68	88.46	92.91
Economics (8.43)	SFAT	248,614	233	204	0.57	55	78.76	87.55
	CFAT	98,106	1,057	1,004	0.42	137	87.99	94.99

* FAT = FA Term, SFAT = single FA Term, CFAT = compound FA Term

column "Irrelevant terms selected" refers to the number of irrelevant FA Terms selected by the system. Hence, precision and recall are calculated as follows:

$$precision = \frac{FA\_Terms\_selected\_automatically}{FA\_Terms\_selected\_automatically + irrelevant\_terms\_selected} \tag{9}$$

$$recall = \frac{FA\_Terms\_selected\_automatically}{Total\_FA\_Terms} \tag{10}$$

5.3 Evaluation of the FA terms dictionary

Once an FA Term is selected, we check if it exists under the same field in the FA Terms dictionary already. If it does not exist under the same field, we compare if it exists under different fields and then we update its level. Then, we append it to the dictionary under the field it is selected for. In the same way, we append all selected FA Terms to the FA Terms dictionary. The resulting FA Terms dictionary looks like Table 5. The column 'Field code' identifies the path of the field to which the FA Term belongs in the field tree. For instance, the field code 14.3.12.1 for the FA Term "Clementine Vulgate" identifies the path <Study/Humanities/Religion/Christianity>. In this experimental evaluation, we constructed a FA Terms dictionary of 31,234 FA Terms for 21 fields.

Table 5 A snapshot of the FA Terms Dictionary

FA Term	Field code	Field Name	Level	*normalized_final_weight*
Clementine Vulgate	14.3.12.1	Christianity	1	0.6193
Cleveland Browns Stadium	13.0	Sports	3	0.4778
Cliff Burton	01.6.0	Music	1	0.5212
Clifford Roberts	13.0	Sports	3	0.4982
Clinton administration	10.0	Politics	3	0.2851
Co-ed	3.0	Education	3	0.5108
Coach	2.19	Transport	4	0.5654
Coach	13.0	Sports	4	0.6654

Table 6 Results of field identification using the FA Terms dictionary

Field	Field code	Wikipedia		RCV1		Newsgroup	
		*n**	*accuracy%*	*n*	*accuracy%*	*n*	*accuracy%*
Sports	13.0	510	100	1,000	100	–	–
Christianity	14.3.12.1	500	100	–	–	997	99.4
War	14.3.1.1	590	100	1,000	91.00	–	–
Politics	10.0	–	–	2,000	98.52	750	85.2
Motor vehicles	6.5	–	–	1,757	99.94	973	99.9
Total *n in ()*, Average *accuracy%*		(1,600)	100	(5,757)	97.90	(2,720)	95.66

* *n* = number of documents

We then tested the quality of the FA Terms dictionary by its ability to identify the fields of 10,077 test documents collected from three different sources: Wikipedia, Reuters RCV1 corpus and 20 Newsgroup data set. The field of a test document is determined by calculating for each field the sum of the product of the term frequency and *normalized_ final_weight* of FA Terms in the document. The term frequency refers to the number of times a particular FA Term occurs in the document, and *normalized_ final_weight* of a FA Term comes from the FA Terms dictionary. The field with the highest sum is identified as the field to which the document belongs. As Table 6 shows, the results are very encouraging. The column *n* stands for the number of test documents and '*accuracy'* stands for the percentage of documents whose fields were correctly identified.

5.4 Discussion of results

A total of 497 to 2,517 FA Terms including both single and compound FA Terms are selected for each field. In total, 22,229 compound FA Terms are selected from 3,700,278 compound FA Term candidates and 9,005 single FA Terms are selected from 8,437,691 single FA Term candidates. This makes the total number of FA Terms selected for the 21 fields 31,234. The relevance of the automatically selected FA Terms was checked by three PhD students. For single FA Terms selection, precision range from 73.76 to 96.07% and recall range from

67.46 to 98.1%. For compound FA Terms selection, precision range from 83.79 to 97.27% and recall range from 65.41 to 96.76%.

The size of domain-specific corpora varied from 2.6 MB for the field <Aviation> to 31.7 MB for <Politics>. The highest number of FA Terms is selected for the field <Motor vehicles> although its corpus size is modest at 6.3 MB, while the lowest number is selected for the field <history> which has a corpus size of 29.3 MB but contains texts from various other fields. We concluded that both the size and the quality of the domain-specific corpus, as well as the choice of a good reference corpus are important factors in extracting a larger number of FA Terms at high precision and recall.

Using the improved traditional method, Sharif et al. [32] reported precision and recall of 98 and 94% respectively. In the new method, we achieved precision of up to 97% and recall up to 98%. However, the two results cannot be compared directly because the experimental setup and environment are different. But the new approach has selected 31,234 FA Terms for just 21 fields compared to the 25,869 FA Terms selected from documents of 850 fields by Sharif et al. [32]. This shows that the new method has selected an average of 1,487 FA Terms per field, while the traditional method selected just around 30 FA Terms per field.

Moreover, the new method is superior because the traditional methods are plagued by the drawbacks described in Sect. 2.4. Traditional methods offer no technique for the automatic extraction and selection of compound FA Terms, although compound FA Terms form a majority of the relevant FA Terms in a given field. Traditional methods use "Concentration Ratio" based only on the term frequency to select FA Terms from FA Term candidates.

The FA Terms dictionary constructed using our method correctly identified all 1,600 test documents taken from Wikipedia while it achieved field identification accuracy of 97.90% with 5,757 test documents of Reuters RCV1 corpus and 95.66% with 2,720 test documents of 20 Newsgroup data set. This high field identification accuracy shows that the quality of FA Terms dictionary constructed by the presented method is high.

6 Conclusion

The novel technique of using FA Terms holds much potential for use in many areas of information retrieval and natural language processing, but one of the major problems today is the lack of a comprehensive FA Terms dictionary. Therefore, we have presented a methodology to extract and select FA Terms effectively to build a comprehensive FA Terms dictionary. The methodology is based on POS pattern rules, corpora comparison and modified *tf-idf* weighting for selecting domain-relevant terms.

Experimental evaluation carried out for 21 different fields using 306 MB of domain-specific corpora obtained from Wikipedia dump selected 22,229 compound FA Terms and 9,005 single FA Terms. The precision and recall were 74–97 and 65–98% respectively. The results show that the proposed methodology is effective for building a comprehensive dictionary of FA Terms.

Future studies will further improve the proposed methodology by adding a document classification module so that documents can be classified automatically and FA Term candidates extracted from them. We will also explore the application of FA Terms in cross language retrieval, domain-specific ontology building and machine translation etc.

References

1. Atlam E, Fuketa M, Morita K, Aoe J (2003) Documents similarity measurement using field association terms. Inf Process Manag 39(6):809–824
2. Atlam E, Ghada E, Morita K, Fuketa M, Aoe J (2006) Automatic building of new field association word candidates using search engine. Inf Process Manag 42(4):951–962
3. Atlam E, Morita K, Fuketa M, Aoe J (2002) A new method for selecting English field association terms of compound words and its knowledge representation. Inf Process Manag 38(6):807–821
4. Bennet NA, He Q, Powell K, Schatz BR (1999) Extracting noun phrases for all of MEDLINE, In: Proceedings of the AMIA symposium. pp 671–675
5. Broughton V (2007) A faceted classification as the basis of a faceted terminology: conversion of a classified structure to thesaurus format in the bliss bibliographic classification, 2nd edn. Axiomathes 18(2):193–210
6. Brunzel M, Spiliopoulou M (2007) Domain relevance on term weighting. Lecture notes in Computer Science, vol 4592. Springer, pp 427–432
7. Collier N, Nobata C, Tsujii J (2002) Automatic acquisition and classification of terminology using a tagged corpus in the molecular biology domain. J Terminol, John Benjamins 7(2):239–257
8. Dozawa T (1999) Innovative multi information dictionary Imidas'99. Annual series. Japan: Zueisha Publication Co. [in Japanese]
9. Drouin P (2004) Detection of domain specific terminology using corpora comparison. In: Proceedings of the 4th international conference on language resources and evaluation (CLREC), pp 79–82
10. Fuketa M, Lee S, Tsuji T, Okada M, Aoe J (2000) A document classification method by using field association words. Int J Inf Sci 126:57–70
11. Graham-Cumming J (2005) Naive Bayesian text classification: fast, accurate, and easy to implement, Dr. Dobb's Journal, http://www.ddj.com/development-tools/184406064, [Accessed 3 Sep 2009]
12. Jiang G, Sato H, Endoh A, Ogasawara K, Sakurai T (2005) Extraction of specific nursing terms using corpora comparison. In: Proceedings of the AMIA annual symposium, 2005:997
13. Jing L, Ng M, Huang J (2009) Knowledge-based vector space model for text clustering, Knowledge and information systems, Springer, London, published online October 2009
14. Jones K (2004) A statistical interpretation of term specificity and its application in retrieval. J Doc 60(5):493–502
15. Krauthammer M, Nenadic G (2004) Term identification in the biomedical literature. J Biomed Inf 37(6):512–526
16. Lan M, Tan C, Low H, Sung S (2005) A comprehensive comparative study on term weighting schemes for text categorization with support vector machines. In: Posters proceedings of 14th international world wide web conference, pp 1032–1033
17. Lee S, Shishibori M, Sumitomo T, Aoe J (2002) Extraction of field-coherent passages. Inf Process Manag 38(2):173–207
18. Leopold E, Kindermann J (2002) Text categorization with support vector machines: how to represent texts in input space? Mach Learn 46(1–3):423–444
19. Lu W, Lin R, Chan Y, Chen K (2008) Using web resources to construct multilingual medical thesaurus for cross-language medical information retrieval. Decis Support Syst 45(3):585–595
20. Nguyen T, Phan T (2007) Using hybrid solution for CLIR noun phrase translation. In: Proceedings of the 9th international conference on information integration and web-based applications & services (iiWAS2007)
21. Pang S, Kasabov N (2009) Encoding and decoding the knowledge of association rules over SVM classification trees. Knowl Inf Syst 19(1):79–105
22. Patry A, Langlais P (2005) Corpus-based terminology extraction. In: Proceedings of the 7th international conference on terminology and knowledge engineering, Copenhagen, Denmark, pp 313–321
23. Peng T, Zuo W, He F (2008) SVM based adaptive learning method for text classification from positive and unlabeled documents. Knowl Inf Syst, Springer, London 16(3):281–301
24. Pinto H, Martins J (2004) Ontologies: how can they be built? Knowl Inf Syst 6(4):441–464
25. Ramakrishnan N (2009) The pervasiveness of data mining and machine learning. Computer 42(8):28–29
26. Rokaya M, Atlam E, Fuketa M, Dorji T, Aoe J (2008) Ranking of field association terms using co-word analysis. Inf Process Manag 44(2):738–755
27. Rose T, Stevenson M, Whitehead M (2002) The reuters corpus Vol. 1- from yesterday's news to tomorrow's language resources. In: Proceedings of the 3rd international conference on language resources and evaluation

28. Salton G, Allan J, Buckley C (1993) Approaches to passage retrieval in full text information systems. In: Proceedings of the 16th annual international ACM/SIGIR conference on research and development in information retrieval, pp 49–58
29. Saneifar H, Bonniol S, Laurent A, Poncelet P, Roche M (2009) Terminology extraction from log files, database and expert systems applications. Lect Notes Comput Sci 5690:769–776
30. Schmid H (1994) Probabilistic part-of-speech tagging using decision trees. In: Proceedings of international conference on new methods in language processing
31. Sclano F, Velardi P (2007) TermExtractor: a web application to learn the shared terminology of emergent web communities. In: Proceedings of the 3rd international conference on interoperability for enterprise software and applications I-ESA 2007
32. Sharif UM, Ghada E, Atlam E, Fuketa M, Morita K, Aoe J (2007) Improvement of building field association term dictionary using passage retrieval. Inf Process Manag 43(2):1793–1807
33. Smadja F (1993) Retrieving collocations form text: xtract. Comput Linguist 19(1):143–177
34. Srinivasan P, Pant G, Menczer F (2005) A general evaluation framework for topical crawlers. Inf Retr 8(3):417–447
35. Tsuji T, Nigazawa H, Okada M, Aoe J (1999) Early field recognition by using field association words. In: Proceedings of the 18th international conference on computer processing of oriental languages, pp 301–304
36. University of Stuttgart, TreeTagger—a language-independent part-of-speech Tagger, http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/ [Downloaded 2 June 2008]
37. Velardi P, Navigli R, D'Amadio P (2008) Mining the web to create specialized glossaries. IEEE Intell Syst 23(5):18–25
38. Voutilamen A (1993) NPtool, a detector of english noun phrases. In: Proceedings of the workshop on very large corpora: academic and industrial perspectives, pp 48–57
39. Wang P, Hu J, Zeng H, Chen Z (2008) Using wikipedia knowledge to improve text classification. Knowl Inf Syst 19(3):265–394
40. Wikipedia Foundation, Inc., English Wikipedia Dumps, http://download.wikimedia.org/enwiki/ [Downloaded 24 July 2008]
41. Wright SE, Budin G (1997) Handbook of terminology management, vol. 1, Basic aspects of terminology management. Amsterdam, Philadelphia, John Benjamins

Author Biographies

Mr. Tshering Cigay Dorji graduated with a B.E. Honors First Class degree in Electrical Engineering from the University of Wollongong, Australia in 1999. He then worked as a System and Database Administrator for Bhutan Telecom prior to coming to Japan to continue his studies under the Monbukagakusho Scholarship program. He received his Masters in Engineering, specializing in Information Science and Intelligent Systems from The University of Tokushima, Japan in 2007. He is currently a Ph.D. student at the same university.

Dr. El-Sayed Atlam received B.Sc. and M.Sc. Degrees in Mathematics from Tanta University, Egypt, in 1990 and 1994 respectively, and Ph.D. in Information Science and Intelligent Systems from the University of Tokushima, Japan in 2002. He received a *Japan Society of the Promotion of Science* (*JSPS*) postdoctoral fellowship from 2003 to 2005 and is currently an assistant professor in Information Science and Intelligent Systems at the University of Tokushima. He is also Associate Professor at the Department of Statistical and Computer Science, Tanta University, Egypt. He is a member of the Computer Algorithm Series of the IEEE Computer Society Press (CAS) and the Egyptian Mathematical Association (EMA). His research interests include information retrieval, natural language processing and document processing.

Dr. Susumu Yata received his B.E., M.E., and Ph.D. degrees in Information Science and Intelligent Systems from the University of Tokushima, Japan, in 2004, 2006, and 2008 respectively. He is currently a Researcher.

Dr. Masao Fuketa is an Associate Professor in Information Science and Intelligent Systems at the University of Tokushima, Japan where he was a Research Assistant from 1998 to 2000. He is a member of the Information Processing Society of Japan and the Association for Natural Language Processing of Japan. He received his B.Sc., M.Sc. and Ph.D. in Information Science and Intelligent Systems from the University of Tokushima in 1993, 1995 and 1998, respectively. His research interests are sentence retrieval from huge text databases and morphological analysis.

Dr. Kazuhiro Morita is an Associate Professor in Information Science and Intelligent Systems at the University of Tokushima where he was a Research Assistant from 2000 to 2006. He is a member of the Information Processing Society of Japan and IEEE Computer Society. He received his B.Sc., M.Sc. and Ph.D. in Information Science and Intelligent Systems from University of Tokushima, Japan, in 1995, 1997 and 2000, respectively. His research interests are sentence retrieval from huge text data bases, double-array structure and binary search tree.

Prof. Jun-ichi Aoe received B.Sc. and M.Sc. in electronic engineering from the University of Tokushima, Japan in 1974 and 1976, respectively, and the Ph.D. in communications engineering from the University of Osaka in 1980. Since 1976 he has been with the University of Tokushima and is currently a Professor in Information Science & Intelligent Systems. His research interests include design of an automatic selection method of key search algorithms based on expert knowledge bases, natural language processing, a shift-search strategy for interleaved LR parsing, robust method for understanding NL interface commands in an intelligent command interpreter, and tried compaction algorithms for large key sets. He is an editor of the Computer Algorithm Series of the IEEE computer Society Press. He is a member of the Association for Computing Machinery, the Association for the Natural Language Processing of Japan and the IEEE Computer Society.