IJCSI International Journal of Computer Science Issues, Vol. 8, Issue 4, No 1, July 2011
ISSN (Online): 1694-0814
www.IJCSI.org

A New Weight Function for Constructing Field Association Terms using Concurrent Words

Elsayed Atlam*

Department of Information Science and Intelligent Systems
University of Tokushima Tokushima, 770-8506, Japan.

Abstract

Field Association (FA) *words* or phrases are serving to identify document fields by reading only some specific words. Document fields can be decided efficiently if there are many rank 1 *FA words* (words that direct connect to terminal fields) and if the *frequency* rate is high. This paper proposes a new method for increasing rank 1 *FA words* using *declinable words* and *concurrent words* which relate to narrow association categories and eliminate *FA word* ambiguity. *Concurrent words* become *Concurrent Field Association Words (CFA words)* if there is a little field overlap. Usually, efficient *CFA words* are difficult to extract using only frequency, so this paper proposes weighting according to *degree of importance* of *concurrent words.* The new weighting method causes *Precision* and *Recall* to be significantly increased by 30% and 40% than by using frequency alone. Moreover, combining *CFA words* with *FA words* allow our new system to append automatically around 28% of *CFA* words to the existence *FA word* Dictionary. Furthermore, Recall is improved by 21% over the recall of the traditional method.

Keywords: *FA Words, Declinable Words, Concurrent Words, CFA words, Recall, Precision.*

1. Introduction

With increasing popularity of the Internet and the tremendous amount of on-line text, automatic classification of document fields [1], *Vector Model* [2], *Probabilistic Model [3-4]* and *Topic Detection* [5-7] can use general information about document fields to calculate degree of document similarity. However, there are problems because of multiple topics and collections of document fields, and so content to be searched usually exists only in part of the file [8-11].

In this paper, *field* means basic and common knowledge used in human communication [12-13]. Readers know topic *super-field* (e.g. *Sports*) or *sub-field* (e.g. *Baseball*) of document fields based on specific *Field Association (FA) words* in that document. For example, the word "*election*" can indicate *super-field* <Politics> and the word "*home run*" can indicate *sub-field* <Baseball>. This novel technique based on *FA words* has been found to be very effective in document classification [13-15], similar file retrieval [16] and passage retrieval [17-18]. This technique also holds much promise for application in many other areas such as domain-specific ontology construction [19], text clustering [20], cross-language retrieval [21], etc.

In this paper, *FA words* are ranked according to document field. Rank 1 *FA words* are relatively few and can be used efficiently to decide document fields. *FA words* in other ranks are always numerous and are not so helpful for deciding document fields. Document fields can be decided easily if there are many rank 1 *FA words* and frequency rate is high. Document fields can not be decided easily if there are few rank 1 *FA words* or if the *FA words* appear in overlapping document fields.

To overcome problems associated with rank 1 *FA words,* this paper proposes a new method using *declinable words* and *concurrent words* to create a r elatively large number of rank 1 *FA words* and to eliminate ambiguous *FA words.*

a) *Declinable words* are words express action, condition or use of things. To eliminate ambiguity of the *FA word, declinable words* are combining with *FA words*. So, for association words that have meanings of variable fields, it is understandable that if we combine the *FA words* together with the *declinable words* which express its action or condition, we can recognize the specific field. Such combination is generally called 'common information'. For example, *FA word* "*pass"* is ambiguous and associated with many sub-fields of <SPORTS>, but combining *declinable word* "*through"* with "*pass*" creates "*through-pass*" which associate with *sub-field* <Soccer>. Such

• Dr. Elsayed Atlam, **Present Address**: Dept. of Information science and Intelligent Systems, Tokushima University, Japan.
Permanent Address: Associate professor at Department of Statistics and Computer Science, Tanta University, Egypt.

method is considered to be possible to limit the *FA words* of association fields

b) *Concurrent words (C words)* usually have two short unit *FA words* connected by particles (e.g. the, in, and) and short unit information can be used to associate with fields. The weight function of *C words* can be expressed by the weight function of the short unit *FA words*.

Section 2 of this paper introduces *FA words* used in this research and describes their construction. Problems related to constructing *FA words* are explained and overcome using *declinable words*. Section 3 describes field association information and *C words,* explaining new weight functions according to *degree of importance* for these *C words.* Section 4 provides simulation results that confirm the efficiency of the weighting method proposed in this research. Section 5 presents a conclusion and indicates possible future work.

2. Field Association Words

Field Association (FA) words can be a w ord (e.g. *game*) or a phrase (e.g. *victory* and *defeat*) that indicates subject matter category in the classification scheme. The basic concept underlying *FA words* involves choosing a limited set of words that best match a given document, so *FA words* describe a set of discriminating words. *FA words* are not always the same as words that specifically identify subject fields. *FA words* appear in a document, but subject words may not appear in that document, so *FA words* may be better for discriminating between documents than subject words [15[18]. Many *FA words* are not subject words (e.g. *case* or *use*). There are few semantic differences among *FA words* and the choice of *FA words* used in a document is mainly a matter of style. *FA words* can identify documents. *Short unit association words* are minimum meaningful units which can not be divided without loss of meaning [18], [22-25]. For example, "*pitcher"* and *"home run"* are short unit *association words* that can be associated with terminal field <*Baseball*>.

2.1 Document Field Tree

A document field *tree structure* represents relationships between ranked document fields [5][6][10][26-28]. In field *tree structure,* a *leaf node* is a *terminal* document field and other nodes are *middle* document fields. In this study, based on Imidas'99 [29] term dictionary, the field tree contains 14 *main (parent) fields*, 18 *middle fields* and 172 *terminal (child) fields*. When there is no conflict *root names* are omitted and only terminal fields are described. For example, in Fig. 1, <*SPORTS*/*Ball Games/ Tennis*> describes document field <*Tennis*> as a *terminal field* of <*Ball Games*>, which is a *middle field* of <*SPORTS*>.

The field *tree structure* classifies document data files. Then, the extraction pattern for common relation is expressed by the number of part of speech which is assigned [12][30] and words are extracted from each field. The frequency rate of extracted words is calculated and *FA words* in each field are decided. Then *FA words* is obtained and registered in the *FA words dictionary*. Words not registered are not taken as *FA words*.

2.2 Ranking *FA Words*

FA words extracted from Corpus data may have various association field ranks. Some *FA words* may associate with only one *terminal field* or one *middle field;* other *FA words* may associate with several *terminal fields* or several *middle fields*. *FA word w* may be defined according to five ranks:

Rank 1: *Complete FA word w* associates with only one *terminal field*.

Rank 2: *Quasi complete FA word w* associates with a limited number of *terminal fields* which have the same *parent field* (*Super-field*).

Rank 3: *Middle FA w* associates with only one *middle field*.

Rank4: *Intersection FA word w* associates with several *middle fields* or several *fields*.

Rank 5: *Non association- word w* does not associate with any specific field.

Table 1 shows some *FA words* of various ranks: (rank 1) "*home run*" and "*Mozart*" associate with only one *terminal field* (<*Baseball*> and <*Classic Music*>*);* (rank 2) "*single match*" and "*paints*" associate with a limited number of *terminal fields* (<*Tennis*>, <*Table Tennis*>; <*Japanese-Style Painting*>, <*Western-Style Painting*>*)* having the same *parent fields* (<SPORTS>; <CULTURE & FINE ARTS>); (rank 3) "*ball"* associates with only one *middle field* <*Ball Game*>*;* (rank 4) "*rule*" associates with several fields (<*SPORTS*>*;* < *Amusement\Game*>*);* (rank 5) "*circumstance*" associates with no field.

Table 1 Paths and Ranks of selected *FA Words*

FA Words	Paths	Ranks
home run	< SPORTS \Ball Game\Baseball>	1
Mozart	<CULTURE & FINE ARTS\Music\Classic Music>	1
single match	< SPORTS \Ball Game\Tennis & Table Tennis>	2
paints	< CULTURE & FINE ARTS\ Art\ Western-Style Painting & Japanese-Style Painting>	2
ball	< SPORTS\Ball Game>	3
rule	< SPORTS >, <Amusement\ Game>	4
circumstance	No Association Field	5

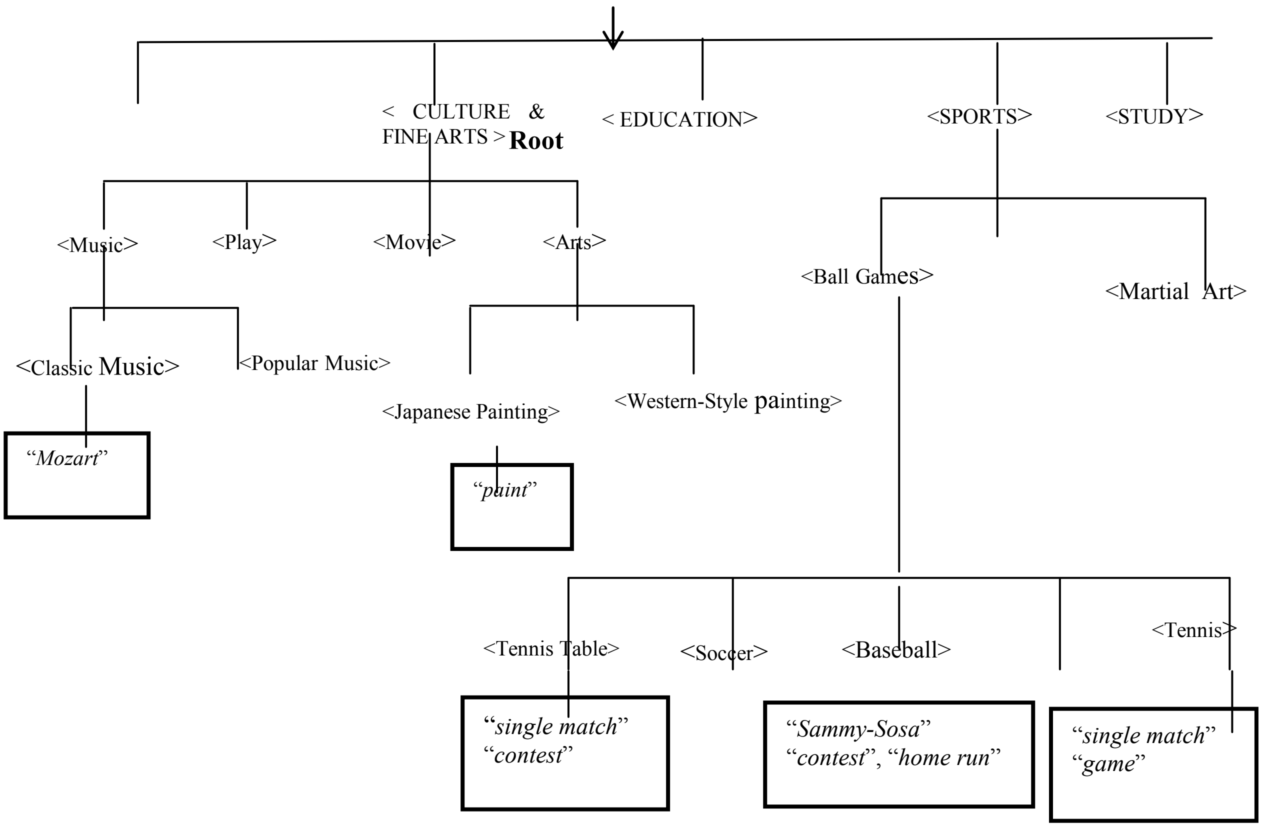


Fig.1 Sample Document Field Tree Highlighting *FA words*

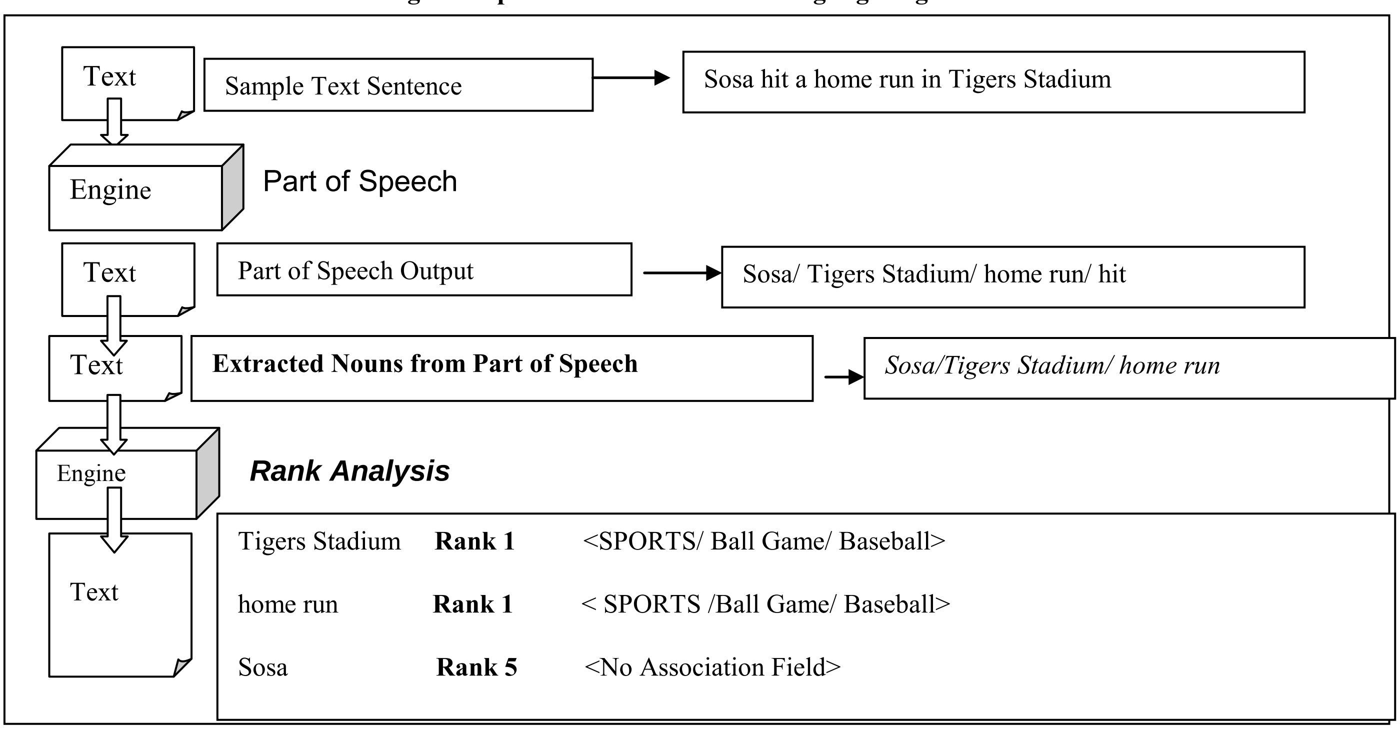


Fig.2 Automatically Constructing *FA Words* and Ranks

IJCSI International Journal of Computer Science Issues, Vol. 8, Issue 4, No 1, July 2011
ISSN (Online): 1694-0814
www.IJCSI.org

2.2 Constructing *FA Words*

2.2.1 Basic Outline

To construct *FA words*, it is first necessary to extract candidate *FA words* from Corpus files classified manually into related fields. Table 2 shows some extracted candidate *FA words* providing information such as: candidates (*A)* are extracted from field (*B)* at a *frequency* (*x times)*.

Table 2 Example of *FA words* candidates

FA words Candidates	Association Fields	Frequency
home run	<Baseball>, <History>	988, 7
paints	<Japanese-Style Painting>, <Western-Style Painting>	10, 8
Virus	<Influenza>, <Cancer>, <Horse Racing>	54, 26, 5
candidate	<Election>, <Congress>,<Judiciary>	395, 38, 3
Griswold	<Literature>, < Western -Style Painting>	12, 2

Fig. 2 shows the main procedure for automatically constructing *FA words* from a Corpus file. The extraction method makes *part of speech* on the Corpus file, extracts nouns from the analysis results and calculates extraction frequency of those nouns. The engine ranks *FA words* using a *field tree*.

2.2.2. The Determination of *FA Words*

This subsection explains the traditional algorithm that automatically determines the candidates for *FA words* and their ranks. In this algorithm, normalized word frequency is used instead of word frequency in each field as follows:

Let *Total_Frequency* (*<T>*) be the total frequency of all words in the terminal field *<T>*; let *Frequency (w, <T>*) be the frequency of the word *w* in the terminal field *<T>*, the normalized frequency (*Normalization (w, <T>*)) can be defined as in the following formula (1):

$$Normalization(w,<T>)=\left(\frac{Frequency(w,<T>)}{Total\_Frequency(<T>)}\right) \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots (1)$$

The normalized frequency defines how much a specific word is concentrated in a specific field.

Definition:

For the parent =*< S>*, the child field = *<C>*, the concentration ratio (Concentration (*w*, *<C>*)) of the *FA word w* in the field *<C>* is defined as in the following formula (2):

$$Concentration(w,<C>)=\left(\frac{Normalization(w,<C>)}{Normalization(w,<S>)}\right) \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots (2)$$

The following algorithm determines *FA words* by considering their ranks.

2.2.3 *FA Words* Determination Algorithm

Input:

(a) *w*, candidates for *FA words,*
(b) Normalization (*w*, <C>) for *w* and for field <C>,
(c) a threshold α, to judge *FA words ranks,*
(d) a field tree.

Output:

associated *FA words* and their ranks for *w.*

(Step 1): Determination of *Complete FA word*

For the root = <S>, the child field = <S/C> of the field tree, the following conditional formula (3) is used to judge whether or not the word *w* is a *Complete FA word.*

$$Concentration(w, <S>) \geq \alpha \ldots\ldots\ldots\ldots\ldots\ldots\ldots (3)$$

If the condition formula (3) is fulfilled, *<S/C>* is replaced by *<S>* and the same judgment is carried out on the field *<S/C>*. By repeating the same determination process, if *<S/C>* becomes a terminal field, *w* is determined as a *Complete FA word* in the field *<S/C>*. If the field *<S/C>* can not fulfill the condition in formula (3), then the process enters (Step 2).

(Step 2): Determination of *Quasi complete* or *Middle FA* words

If *w* is not determined as a *Complete FA word* in the field *< S/C >*, the terminal field has not been reached. Therefore, the field *<S>* should be a medium field and has at least two or more (*m*≥2) children fields. From all children fields $<S/C_k>$ (*1<k<m*) of the medium field *<S>* calculate the average value of *k* times children including word *w* as in the following formula (4):

$$\left(\frac{\sum_{k=1}^{m} Normalization(w, <C_k>)}{m}\right) \ldots\ldots\ldots\ldots\ldots\ldots\ldots (4)$$

Accumulated Concentration (*w*, $<S/C_k>$) ratio for the children fields has higher normalized frequencies than the average value in formula (4). If the accumulated concentration ratio of *k* times (*1<k<m*) exceeds α and the children fields $<S/C_k>$ are all terminal fields, *w* is judged as a *Quasi complete FA word* in fields $<S/C_k>$. If the accumulated value does not exceed the threshold α, *w* is determined as a *Middle FA word* of field *<S>*. However, if all of these children fields are not terminal fields, the process enters (Step 3) and conducts the determination process of *Intersection FA word*.

(Step 3): Determination of *Intersection FA word*

Extract the terminal field <*S/C*> from *k* children fields and determine *w* as a *Intersection FA word* of the field < *S/C*>. Except for the terminal fields the child field <*S/C*> is changed into root <S> of the field tree, repeat the process to conduct (Step 1) and (Step 2). Then, many medium fields and terminal fields are obtained, and *w* is judged as a *Multiple FA word* of the field <*S*>.

(***end of algorithm)***

2.2.4 *FA Words* and *Declinable Words*

To eliminate ambiguity, the present method combines *FA words* with *declinable words* from Corpus documents which are classified beforehand by *Tree* structure.

In Table 3, "*pass"* is an *FA word* of rank 2 for <Baseball>, <Soccer>, and <Basketball>. However, in this document data the action "*through-pass*" only exists in <Soccer >, so "*through-pass*" is considered to be an *FA word* for <Soccer>. *"Game"* is mainly an *FA word* of rank 3 used in middle field <Ball Game>. However, the action "*a perfect game*" is only in <Baseball >, so "*a perfect game*" is considered to be an *FA word* for <Baseball>. *"Recommendation"* is an *FA word* of rank 4 for fields <Election> and <Entrance Examination>, but "*recommendation recruitment*" only exists in <Entrance Examination>, so "*recommendation recruitment*" is an *FA word* for <Entrance Examination>. "*Fish*" is not an *FA word* because "*fish*" can not be used to associate with any field and could not specify the related filed in our data. However, the condition or the co-occurrence "*Extinction of fish*" is an idiomatic expression used only in a specific field, and so "*Extinction of fish"* is an *FA word* for <Environment Problems>.

As *FA words* form a limited set of words or compound words that form the essence of the field to which they belong. In other words, the *FA words* store the knowledge of the field. Just as humans with prior experience and knowledge can identify the field to which a text belongs. Moreover, for all *FA words*/*declinable words*, fields can not be necessarily specified. For example, "*strike"*, *"hit"* and *"shoot*" have different meaning in <Baseball>, <Soccer> and <Basketball>. Combining "*strike"* or *"hit*" with "*shoot*" does not produce an expression which can be used in any of the three fields. Generally, association fields are not necessary identified by combining *FA words* with *declinable words*. However, the range of association fields can be limited by pairing *FA words* having meaningful relationship. So, this algorithm will explained in detail in the next section 3.1.2.

3. *Concurrent words* and Attaching Weight

3.1.1 *Concurrent Words*

Concurrent words (C words) are two short unit *FA words* connected by particles (e.g. *the, in, and*) which are used to associate fields. The importance of *C words* can be expressed by ranking the weight of the short unit *FA words*. The importance of *C words* relates especially to appearance frequency and to association fields of the short unit *FA words*. The frequency of short unit words shows field rank, and number of overlapping fields shows the degree of ambiguity of the short unit words.

In this paper, it is assumed that no rank 1 short unit *FA words* are *C words* because rank 1 *FA words* refer to specific fields and it is not necessary to converge association fields.

3.1.2 Attaching Weight

Generally, to extract a word which characterizes a file, a weight function *TF x IDF* attaches to the words (*TF* is a high frequency of the appearance characteristic words and *IDF* is inverse document Frequency) [31, 32]. However, not every word with high frequency characterizes a file. For example, particles (the, to, etc) appear often in a file, but the particles are not characteristic words. On the other hand, some characteristic words have relatively low frequency, so *IDF* attaches high weight to those characteristic words [33] and considers weight in many fields. *IDF* value is given by log *N/df(t),* where total number of files is *N* and the number of files which include word *t* is *df(t)*. $TF \times IDF$ is given by:

$$W(d,t) = TF(d,t) \times IDF(t) \cdots\cdots\cdots\cdots\cdots (1)$$

where *TF* is the normalized frequency value of a word *t* in a file *d*.

This research applies $TF \times IDF$ to consider the normalized frequency of a word α in one field *A*. So, the weight of a short unit word α can be defined:

$$Weight_A(\alpha) = Freq_A(\alpha) \times \log(\frac{N}{Category\_num(\alpha)}) \cdots\cdots\cdots\cdots (2)$$

where *Freq* is the normalized frequency of word α in field *A*, *N* is total number of fields and Category_num is number of fields containing α.

If in field *A*, short unit word α appears *100 times,* then α is considered to have strong field association in field *A*. However, if α appears in *100 fields*, then α is judged to be ambiguous. If α appears in only two or three fields, then α is judged to have strong field association, because of high frequency and limited number of associated fields.

In the same way, the weight of word β in Field *A* can be calculated:

$$Weight_A(\beta) = Freq_A(\beta) \times \log(\frac{N}{Category\_num(\beta)}) \cdots\cdots\cdots\cdots (3)$$

Consider a *C word* $\alpha + \beta$ is in a field *A*, the weight of the *C words* is:

Table 3 Sample of *FA Words*/ *Declinable Words* and *Association Fields* with Ranks

FA Words	*Association Fields*	Ranks	***FA Words / Declinable Words***	***Association Fields***	Ranks
pass	<Baseball>, <Soccer>, <Basketball>	2	*through-pass*	< Soccer >	1
game	<SPORTS>	3	*a perfect game*	<Baseball >	1
recommandation	<Election>, <Entrance Examination>	4	*recommendation recruitment*	<Entrance Examination>	1
fish	No Association Field	5	*extinction of fish*	<Environment Problem>	1

$$Weight_A(\alpha+\beta) = Weight_A(\alpha) + Weight_A(\beta) = Freq_A(\alpha)\times\log(\frac{N}{Category\_num(\alpha)}) + Freq_A(\beta)\times\log(\frac{N}{Category\_num(\beta)}) \cdots\cdots(4)$$

Combining weights of α + β allows balance of total weight. If one word has heavy weight and another has light weight, the sum will be heavy. If short unit word α has a heavy weight and association field is decided, after attaching β the association field of *C word* α + β can converge to a specific field. Weight of *C word* α + β is based on the weight of α and when β is attached, the weight of β increases the total weight.

C words are *CFA words* (*Concurrent Field Association Words*) in a limited number of document fields when there is a little field overlap. When *C words* exist in several overlapping fields, they are ambiguous. By calculating weight according to *degree of importance* of *C words* α + β in fields, it requires consideration of ambiguity of each *C word*. In Fig. 3, words α and β exist in fields *D* and *F* at the same time*,* so α and β can be considered *C words* in those fields.

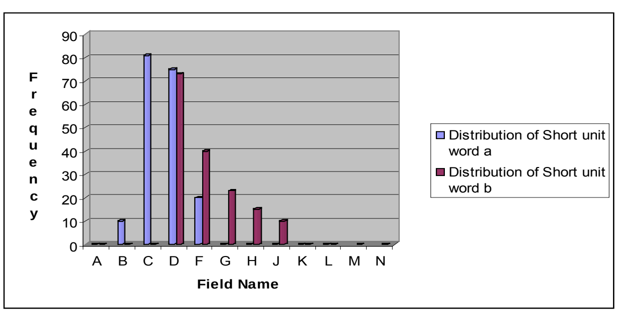


Fig. 3 Field distribution of short unit words α and β

To overcome ambiguity, when α +β occur in many overlapping fields, a new weight function $Weight_{Cross}(\alpha+\beta)$ called *Weight Cross_Category function* is defined:

$$Weight_{Cross}(\alpha+\beta) = \frac{Weight_A(\alpha+\beta)}{Cross\_Category\_num} = \frac{Weight_A(\alpha)+Weight_A(\beta)}{Cross\_Category\_num} = \frac{Freq_A(\alpha)\times\log(\frac{N}{Category\_num(\alpha)}) + Freq_A(\beta)\times\log(\frac{N}{Category\_num(\beta)})}{Cross\_Category\_num} \ldots\ldots(5)$$

where Cross_Category_num is the number of overlapping fields of α and β.

In formula 5, frequency of *C words* is not expressed. Ideally, effective *C words* can be obtained by attaching weight without considering frequency. But frequency must be considered and a new weight according to *degree of importance* of C words is calculated:

$$W_{new}(\alpha+\beta) = Concurrent\_Num \times Weight_{Cross}(\alpha+\beta) \cdots\cdots\cdots(6)$$

where Concurrent _Num is the frequency of the *C word*.

Calculating weight according to *degree of importance* by formula 6 a llows *C words* to be transferred efficiently to *CFA words*.

The following cases are examples of weight according to *degree of importance* of *C words*:

Case (1): *C words* with high frequency are confirmed to be improper for use as *CFA words.*

In field <Soccer>

foreigner Freq. = 52 Category _ num (*foreigner*) = 35
athlete Freq. = 535 Category _ num (*athlete*) = 57
foreigner and *athlete* Freq. = 52 Cross_Category_num = 26
foreigner and *athlete* (*frequency* rank)= 13
foreigner and *athlete* (weighting according to *degree of importance* rank) = 408

$$W_{new}(\textit{foreigner and athlete}) = 52 \times \frac{52 \times \log(172/35) + 535 \times \log(172/57)}{26} = 585.129$$

In field <Soccer>, the concurrent relation of "*foreigner*" and "*athlete*" has frequency of 52. If *C words* are ranked according to *frequency,* provide relatively high rank of 13 in field <Soccer>. So, *C words* might appear to be important by considering only frequency, but the concurrent relation of "*foreigner*" and "*athlete*" is not characteristic words in field *<Soccer>;* "*foreigner*" and "*athlete*" appear in all sub- fields of field *<SPORTS>*.

Ranking "*foreigner*" and "*athlete*" by weighting according to *degree of importance* provides a relatively low rank of 408. So, *C words* "*foreigner*" and "*athlete*" are not *CFA words* in field <Soccer>.

Case (2): *C Words* with low frequency are confirmed as *CFA words*.

In field <Soccer>

loop Freq. = 8 Category _ num ("*loop*") = 2
shoot Freq. = 284 Category _ num ("*shoot*") = 6
loop and *shoot* Freq. = 8
Cross_Category_num = 1
loop and *shoot* (*frequency* rank) = 106
loop and *shoot* (weighting according to *degree of importance* rank) = 15

$$W_{new}(\textit{loop and shoot}) = 8 \times \frac{8 \times \log(172/2) + 284 \times \log(172/6)}{1} = 3435$$

The frequency of *C words* "*loop* and *shoot*" has frequency of 8 with relatively low rank of 106 compared to "*foreigner* and *athlete*". However, "*loop* and *shoot*" can be considered as *CFA words* in field *<Soccer>*.

Ranking "*loop*" and "*shoot*" by weighting according to *degree of importance* provides a relatively high rank of 15. So, "*loop*" and "*shoot*" are *CFA words* for <Soccer>.

Case (3): Both *C words* are ambiguous, but they can become *CFA words* by combining them.

In field <Soccer>

goal Freq. = 406 Category _ num ("*goal*") = 17
upper left Freq. = 2 Category _ num ("*upper left*") = 5
goal of *upper left* Freq. = 2
Cross_Category_num = 1
goal of *upper left* (*frequency* rank) = 1447
goal of *upper left* (weighting according to *degree of importance* rank) = 173

$$W_{new}(\textit{goal of upper left}) = 2 \times \frac{406 \times \log(172/17) + 2 \times \log(172/5)}{1} = 722.266$$

"*goal*" in field *<SPORTS>* is an ambiguous *FA word* which overlaps many document fields and the term "*upper left*" does not identify any particular thing. Combining "*goal*" with "*upper left*" identifies specific association sub-field <Soccer>. Ranking "*goal*" and "*upper left*" according to *frequency* provides relatively low rank of 1447, so these terms are not *CFA words* of field *<Soccer>* just because of frequency. Ranking "*goal*" and "*upper left*" by weighting according to *degree of importance* provides relatively high rank of 173, suggesting that those words are *CFA words*.

In brief, we can say that ranked some words according to *frequency,* provide relatively high rank in some fields and might appear to be important by considering only frequency, even it is not true. Ideally, effective *C words* can be obtained by attaching new weight according to *degree of importance* with considering frequency too.

Table 4 *C words* arranged by weighting according to *Degree of Importance*

C Words	*Weighting According to Degree of Importance*	*Frequency*
through – pass	2393.325	66
middle shoot	2169.166	38
World Cup	1399.5	27
coach Sammy McIlory	970.667	66
chairman Whitey Ford	775.238	37
direct pass	762.562	21
pass and join (continue)	632.98	41
France team	499	21
match with Oman team	489	12
loop shoot	410	8
Zinedine Zidane	393	116
pass and transfer	320	21
Long and short pass	303	5
right side	109.524	86
foreign athlete	23.514	52

3.2 *CFA Words and Ranks*

3.2.1 Constructing *CFA Words*

CFA words can be created by attaching weight to *C words* according to *degree of importance* to. *C words* associate more efficiently with fields when there is high weighting according to *degree of importance*. So, it is possible to extract only high ranking of *C words.*

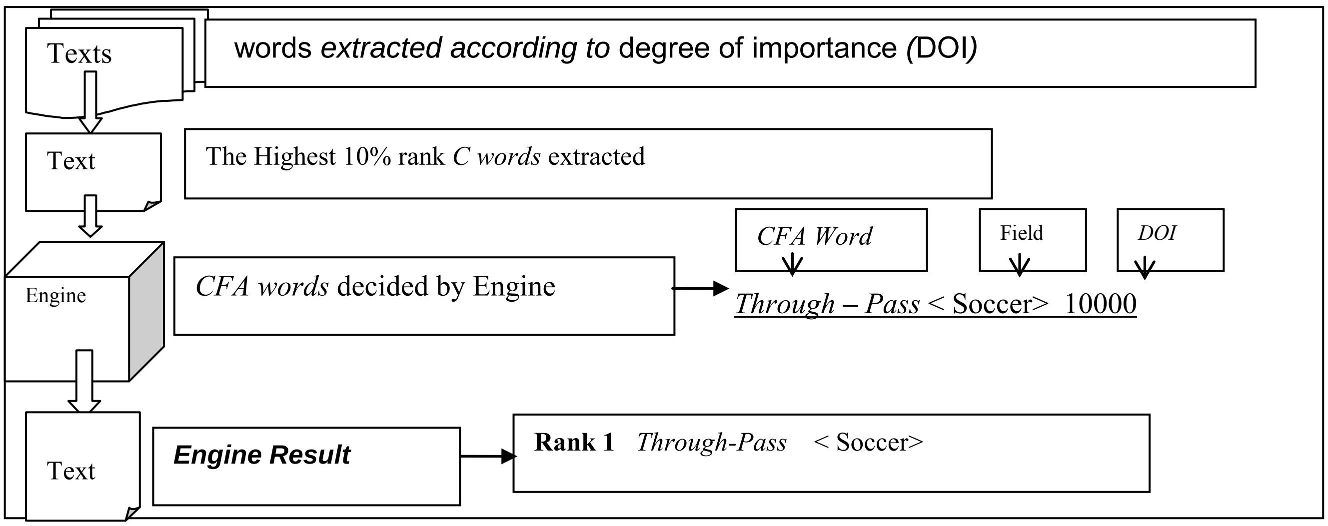


Fig. 4 Automatic construction of sample *CFA words* "*Through-Pass*"

Table 5 *CFA* Words Ranking

Rank	*CFA Words*	*Association Fields*
1	*Organ tone*	<Classic Music>
1	*iron club (tick away)*	<Golf>
1	*right foot*	<Soccer>
1	*Dubai inner tracks*	<Horse Racing>
1	*radioactivity pollution*	<Nuclear Power>
1	*budget vote*	<National Assembly or Congress>
1	*nuclear test prohibit*	<International Law>
1	*blackmail suspicion*	<Judiciary>
1	*approve recognition candidate*	<Election>
1	*the outside lower*	<Baseball>
1	*King era*	<History>
2	*doubles woman*	{<Table Tennis>,<Tennis>}, or {<Game of Go>, <Japanese Chess>}
2	*professional chess player*	<Game of Go>, <Japanese Chess>
2	*husband and wife (different family name)*	< Judiciary>, <Congress>
2	*think for a long time into*	<Game of go>, <Japanese Chess>
2	*shoot and lose target*	<Soccer>, <Basketball>
2	*movie cameraman*	<Film>, <Photo>
3	*Calgary Olympics*	<Winter Sports>
4	*world championship*	<Judo>,<Ski>,<Skate>

3.2.2 *CFA Words Ranks*

Ranks of *CFA words* are decided in the same way that ranks of *FA words* are decided, using the algorithm in section 2.2. Ranks of *FA words* are decided according to *frequency* of words in each document field. However, ranks of *CFA words* are decided by weighting according to *degree of importance*. Table 5 shows examples of *CFA word* rank.

Extracting *C words* of the top 10% rank allows extraction of many words that can be used to determine specific fields; many *C words* are ranked 1 and few are ranked 2, 3 or 4. However, because only extracted *CFA words* are considered in their fields, even *CFA words* of ranks 2, 3 or 4 can be used to determine correct field with little or no manual revision. On the contrary, increasing the number of *C words* increases the number of *CFA words* of ranks 2, 3 or 4, so *Precision* of deciding field decreases.

4. Experimental Evaluation

4.1 Field systems and test data

To verify the efficiency of the new method described in this paper, about 38,000 articles from a data set of 20 Newsgroups from *CNN Web Site* (1995-2001) were selected. There were various topics related to *sports, computers, politics, economics*, etc. This Method is also applied on the large Penn-Treebank English Corpus (Treebank Project Release 2 (1995) [34].

The accumulating method is to search titles of articles by using keywords exists in field tree system. Then, if the keyword is found, the document is classified into the field containing the key word. Roughly classified fields are confirmed manually to extract *C words.*

4.2 Method Evaluation

Precision and *Recall* are evaluated to show how well weighting according to *degree of importance* calculated by the new method expresses *CFA words* in specific fields. *C words* with higher weighting according to *degree of importance* are judged to determine *CFA words*. The highest 10%, 20%, and 30% ranking of weight according to *degree of importance* is used to calculate *Precision* and *Recall.* Efficiency of this method is also estimated.

The test is done by the following sequence:

Step 1: Roughly classified fields are confirmed manually to extract *C words.*
Step 2: Weighting according to *degree of importance* is attached to *C words.*
Step 3: *Precision (P)* and *Recall (R)* are evaluated [33] as follows:

Number of Relevant *CFA words* in the extracted *C words*
Precision (P) = --
Total number of *C words* automatically extracted

Number of Relevant *CFA words* in the extracted *C words*
Recall (R)= --
Total Number of *CFA words* automatically extracted

4.3 Experimentation

Using the method explained in section 3.2.1., selected fields and the number of *C words* for testing are <Soccer, 13191>, <Japanese Chess, 2305>, <Popular Music, 2052>, <Horse Racing, 5645>, <Tennis, 10190>, <Baseball, 15281>, <National Assembly or Congress>, 4172> and <Election, 11875>.

Tables 7 and 8 show *P* and *R* of *C words* in field <Soccer> are arranged by weighting according to *degree of importance* and *frequency*.

Table 7 Precision & Recall *in Field <Soccer> by Weighting According to* Degree o f Importance

Rank	*P* (%)	*R* (%
Upper 10%	79	71
Upper 20%	48	87
Upper 30%	34	93
Upper 40%	26.4	95.7
Upper 50%	21.3	96.7
Upper 60%	18	98.3
Upper 70%	15.6	99

Table 8 Precision & Recall *in Field <Soccer> According to* Frequency

Rank	*P* (%)	*R* (%)
Upper 10%	28	25
Upper 20%	22	40
Upper 30%	30	81
Upper 40%	25	93
Upper 50%	21	96
Upper 60%	18	97
Upper 70%	15	99

Fig. 5 shows the change in *P* and *R* by weighting according to *degree of importance*. *C words* in the highest percentage rank have the highest *P* and *R*. *P* is high because many *CFA words* are in the extracted *C words; R* is high because the total number of *CFA* includes many *CFA words*.

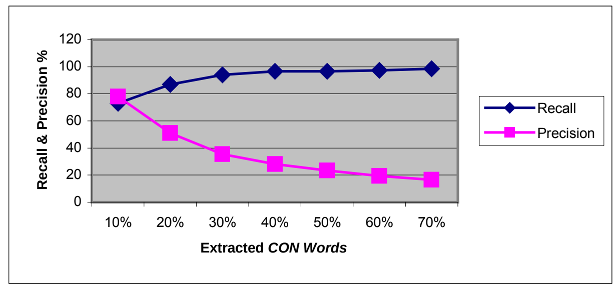


Fig. 5 *Precision* and *Recall* in Field <Soccer> According to *Degree of Importance*

Fig. 6 shows the change in *P* and *R* according to *frequency*. *P* is low even if the number of extracted *C words* increases, because *frequency* alone does not indicate if words are important to document fields. *R* increases with increase in extracted *C words.* When there are few extracted *C words, R* is low compared with *C words* weighted according to *degree of importance*. When arranged according to *frequency*, *C words* are not characteristic in the document fields just because of high *frequency* and there are a few *CFA words* of high ranking.

In <Soccer>, when the rank of extracted *CFA words* increases from 10% to 20%, *P* decreases significantly, showing few *C words* associated with the field. But field <Soccer> has many rank 1 *CFA words* and few rank 2 or rank 3 *CFA words,* causing *P* & *R* to increase.

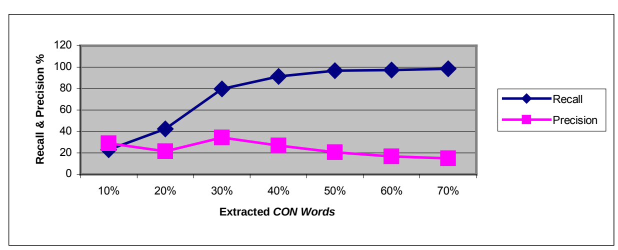


Fig. 6 *Precision* and *Recall* in <Soccer> According *frequency*

In Fig. 7, *P* & *R* of field <Election> are lower than *P* and

R in field <Soccer> weighted according to *degree of importance* because field <Election> has many *CFA words* overlapping many association fields. For example, *C words* classified as *CFA words* for field <Election> are also in sub-field <the Diet> of field <Politics>. Many *C words* can be detected, but fields can not always be decided. For example, only 16 of 64 rank 2 *CFA words* associate with fields <Election> and <the Diet>[*]. Therefore, there are many ambiguous *CFA words* in fields <Election> and <the Diet>, making *P* and *R* lower than in <Soccer>.

Arranged by *frequency*, *P* is similar in fields <Election> and <Soccer> as in Fig. 6 & Fig. 8.

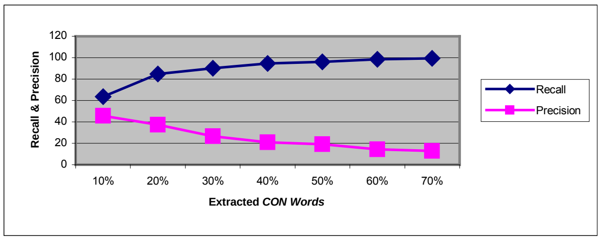


Fig. 7 *Precision* and *Recall* in Field <Election> According to *Degree of Importance*

In Fig. 7 and Fig. 8, *R* is lower when weighting uses *frequency* instead of *degree of importance*. However, *R* values over 40% are relatively similar using *frequency* and weighting according to *degree of importance* because field <Election> has many ambiguous *CFA words.* It is difficult to determine *CFA words,* so *R* does not change even by weighting according to *degree of importance*.

Parent field <Election> has many words associated with child field <the Diet>, so few *CFA words* characterize parent field <Election>. Therefore, field <Election> has many ambiguous *CFA words* which are not useful for deciding fields, causing *P* and *R* to decrease.

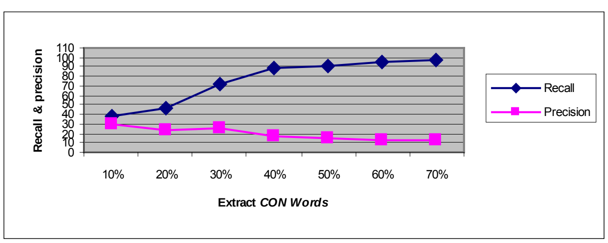


Fig. 8 *Precision* and *Recall* in Field <Election> According to *Frequency*

- *The Diet means in this corpus:*
  1) A national or local legislative assembly in certain countries, such as Japan, or
  2) A formal general assembly of the princes of the Holy Roman Empire.

Fig. 9 and Fig.10 show average *P* and *R* in fields <Selection, Soccer, Popular Music, Horse Racing, Japanese Chess> according to *degree of importance* of *C words*. *P* is 40% higher and *R* is 30% higher than by arranging *C words* according to *frequency*.

frequency.

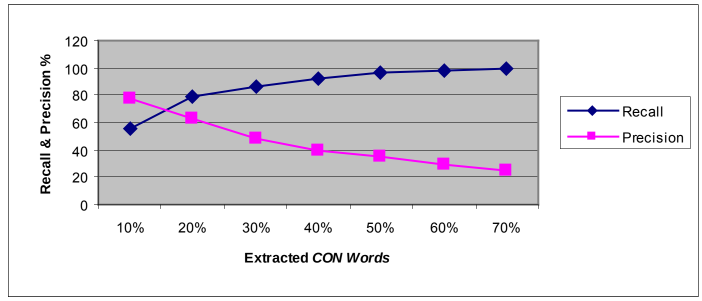


Fig. 9 *Precision* and *Recall* in five Selected Fields according to Degree of *Importance*

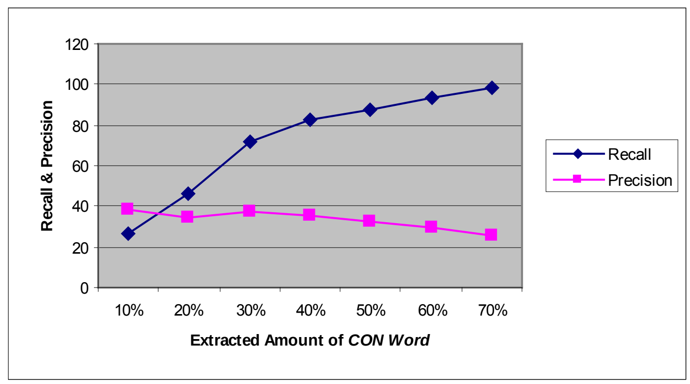
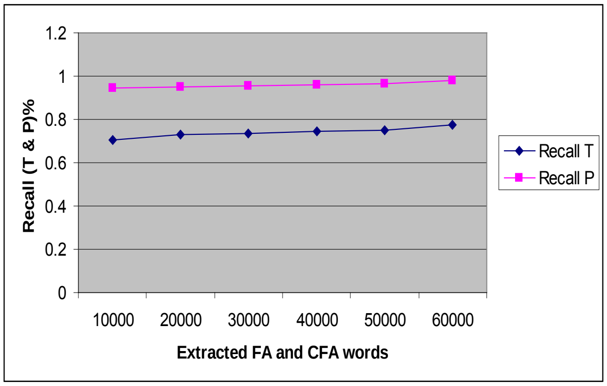


Fig. 10 *Precision* and *Recall* in Five Selected Fields According to *Frequency*

It concludes that *CFA words* can be extracted efficiently if weighting according to *degree of importance* is attached. But, if words are arranged according to *frequency*, *P* seems to be constant, meaning that *CFA words* do not depend on *frequency* alone.

4.4 Recall Improvement

The following part shows the behavior of Recall with extracted *CFA words* after using the traditional [16] and presented methods. Figure 11 shows the effectiveness of appending CFW words on Recall. Using the traditional method reported recall of 77%. In the new method, we achieved recall up to 98%. This means that Recall is improving by 21% after appending more *CFA words* to the existence Dictionary.



Recall T= Recall for Traditional method, Recall P= Recall for the presented method

Figure 11 Recall using the traditional and presented methods

According to Figure 11, it is clear that the Recall of the presented method is improved by 21% higher than the Recall of the traditional method. This is because after appending more *C words*, the numbers of extracted *CFA words* are increase as well as Recall of the presented method.

In conclusion, the presented method performed better Recall than the traditional method

6. Conclusion

Document fields can be decided efficiently if there are many rank 1 *FA words* and if the *frequency* rate is high, but generally, there is limited rank 1 *FA words*, especially when there are few Corpus documents. This paper proposes a method for deciding *FA words* using *C words* and *declinable words* which relate to narrow association categories and eliminate *FA word* ambiguity. Usually, efficient *CFA words* are difficult to extract using *frequency* only. This paper proposes a n ew efficient method for weighting according to *degree of importance* of *C words,* causing *P* and *R* to be higher than by using *frequency* alone. *R* and *P* significantly increase by using *C words* ranked in the top 10% weighted according to *degree of importance. R* and *P* decrease somewhat when *C words* are ranked between 10% to 50% by weighting according to *degree of importance* because there are many ambiguous words. Moreover, combining *CFA words* with *FA words* allow our new system to append automatically around 28% of *CFA* words to the existence *FA word* Dictionary. Furthermore, Recall has been improved by 21% over the recall of the traditional method.

Future research could focus on clustering *C words* and *FA words*. Moreover, we can apply same approach in other languages such as Arabic, French and Chinese.

[1] F. Fukumoto, Y. Suzuki. "Automatic Clustering of Articles using Dictionary definitions". In proceeding of the 16th International Conference on C omputional Linguistic (COLING'96), 1996, pp. 406-411,
[2] L. Jing, J. Huang. "Knowledge-based vector space model for text clustering", Knowledge and Information Systems, Springer London, published online October 2009.
[3] N. Fuhr. "Models for retrieval with probabilistic indexing, Information Processing and Retrieval **25 (1),** (1989) 55-72.
[4] K. Jones. "A statistical interpretation of term specificity and its application in retrieval", Journal of documentation, 60(5), (2004) 493-502.
[5] M. Rogati and Y. Yang, "Resource Selection for Domain Specific CLIR", *ACM SIGIR*, 2004.
[6] K. Salomatin, Y. Yang and A. Lad. Multi-field Correlated Topic Modeling, SIAM International Conference on Data Mining (SDM09), (2009), pp 628-637.
[7] Y. Yang, J. Zhang, J. Carbonell and C. Jin. "Topic-conditioned Novelty Detection", ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, (2002), pp 688-693.
[8] M. Alfio, M. G liozzo, P. Marco and P. Patrick. "The Domain Restriction Hypothesis: Relating Term Similarity and Semantic Consistency". In Proceedings of North American Association for Computational Linguistics/ Human Language Technology (NAACL HLT 07). (2007), pp. 131-138.
[9] L. Breiman, J.H. Friedman, R. A. Olshen and C.J. Stone. *Classification and Regression Trees*. Chapman & Hall, 1984.
[10] G., Elmarhomy, E.-S. Atlam, K., Morita, M. Fuketa and J. Aoe. "Automatic Deletion of Unnecessary Field Association Word Using Morphological Analysis", International Journal of Computer and Mathematics, **83(3),** (2006) 247-262.
[11] G. Jiang, H. Sato, A. Endoh, K. Ogasawara and T. Sakurai. "Extraction of Specific Nursing Terms Using Corpora Comparison", In *Proceedings of the AMIA Annual Symposium*, (2005), pp. 997.
[12] K. Kawabe and Y. Matsumoto, "Acquisition of normal lexical knowledge based on basic level category". Information Processing Society of Japan, SIG note, NL125-9, (1998), pp.87-92 (in Japanese).
[13] T. Tsuji, H. Nigazawa, M. Okada, and J. Aoe. "Early Field Recognition by Using Field Association Words". In the Proceeding of the 18th International Conference on C omputer Processing of Oriental Language, 2, (1999), pp. 301-304.
[14] M. Fuketa, S. Lee, T. Tsuji, M. Okada and J. Aoe. "A Document Classification Method by using Field Association Words", International Journal of Information Sciences, **26,** (2000) 57-70.
[15] M. Rokaya, E.-S. Atlam, K., Morita, M., Fuketa, C. Dorji and J. Aoe, "Ranking of field association terms using Co-word analysis", Information Processing & Management Journal, **44**, (2008) 738-755.
[16] E.-S. Atlam, M., Fuketa, K., Morita and J. Aoe. "Documents Similarity Measurement using Field Association Terms", Information Processing & Management, **39(6)**, (2003) 809-824.
[17] S. Lee, M. Shishibori, T. Sumitomo and J. Aoe. "Extraction of Field-coherent Passages", Information Processing & Management, **38(2),** (2002)173-207.

[18] MD. Sherif Uddin, G., Elmarhomy, E.-S. Atlam, K., Morita, M. Fuketa and J. Aoe, "Improvement of Automatic Building Field Association Term Dictionary Using Passage Retrieval", Information Processing & Management Journal. **143**, (2007) 1793-1807.

[19] H. Pinto J. and Martins. "Ontologies: How can They be Built?", Knowledge and Information Systems, **6 (4),** (2004) 441-464.

[20] K. Jones. "A statistical interpretation of term specificity and its application in retrieval", Journal of documentation, **60(5),** (2004) 493-502.

[21] W. Lu, R. Lin, Y. Chan and K. Chen. "Using Web Resources to Construct Multilingual Medical Thesaurus for Cross-language Medical Information Retrieval", Decision Support Systems, **45(3),** (2008) 585-595.

[22] E.-S. Atlam, K., Morita and J. Aoe. "A New Method for Selecting English Compound Terms and its Knowledge Representation". Information Processing & Management Journal, **38(6)**, (2002) 807-821.

[23] E.-S. Atlam, K., Morita, M. Fuketa and Jun-ich Aoe. "Automatic Building of New Field Association Word Candidates Using Search Engine", Information Processing & Management Journal, **42(4),** (2006) 951-962.

[24] E.-S. Atlam, K., Morita, M. Fuketa and J. Aoe. "A new method using declinable words and concurrent words to construct a large number of FA words", 7th WSEAS International Conference on Computational Intelligence Man-Machine Systems and CYBERNETS (CIMMCS'08), Cairo, Egypt, December 29-31.

[25] T. Tsuji, M. Fuketa, K. Morita, and J. Aoe. "An Efficient Method of Determining Field Association Terms of Compound Words". Journal of Natural Language Processing. 7(2), (2000) 3-26.

[26] J. Aoe, K., Morita, and H. Mochizuki. " An Efficient Retrieval Algorithm of Collocate Information Using Tree Structure". Transaction of The IPSJ, 39 (9),(1989) 2563-2571.

[27] L. Breiman, J.H. Friedman, R. A. Olshen and C.J. Stone. *Classification and Regression Trees*. Chapman & Hall, 1984.

[28] M. Melucii. "Passage Retrieval and a Probabilistic Technique" Information Processing and Management.Vol.34, No.1, pp.43-68, 1998.

[29] T. Dozawa. Innovative Multi Information Dictionary Imidas'99, Annual Series, Zueisha Publication Co., Japan 1999 (In Japanese).

[30] A. Ratnaparkhi, J. Reynar, and S. Roukos. "A Maximum Entropy Model for Prepositional Phrase Attachment", In *Proceedings of the Human Language Technology Workshop* (ARP, 1994), (1994), pp. 250-255.

[31] G. Salton. *Automatic Text Processing: The Transformation, Analysis, and Retrieval of Information by Computer*. Addison-Wesley, 1988.

[32] G. Salton and C.S. Yang. "On the specification of term values in automatic indexing". Journal of Documentation, **29(4),** (1973) 351-372.

[33] G .Salton, & M.J. McGill, *Introduction of Modern Information Retrieval*. New York: McGraw-Hill,1983.

[34] Treebank Project Release 2, *1 million words of 1989 Wall Street Journal material annotated in Treebank II style*, University of Pennsylvania, 1995.