NEW METHOD USING DECLINABLE WORDS AND CONCURRENT WORDS TO CREATE A LARGE NUMBER OF FA WORDS

El-Sayed Atlam, K. Morita, M. Fuketa and Jun-ichi Aoe
Department of Information Science and Intelligent Systems, University of Tokushima, Tokushima, 770-8506, Japan

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

Abstract: The Readers can know the subject of many document fields by reading only some specific *Field Association (FA) 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 higher than by using frequency alone. Moreover, combining *CFA words* with *FA words* allow easy search of fields which can not be searched by using only *FA words*.

1 INTRODUCTION

Determining keywords is crucial in successful *Information Retrieval* (*IR*). After automatic classification of document fields (Fukumoto, 1996), *Vector Model* (Iwayama, 1999) and *Probabilistic Model* (Fuhr, 1989) can use general information about document files 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 (Callen , 1994; Melucii, 1998).

In this paper, *field* means basic and common knowledge used in human communication (Tsuji et al., 1999). 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>.

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 relatively large number of rank 1 *FA words* and to eliminate ambiguous *FA words.*

a) *Declinable words* express action or condition. To eliminate ambiguity of the *FA word, declinable words* are combining with *FA words*. 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>.

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*.

Atlam E., Morita K., Fuketa M. and Aoe J. (2010).
NEW METHOD USING DECLINABLE WORDS AND CONCURRENT WORDS TO CREATE A LARGE NUMBER OF FA WORDS.
In *Proceedings of the 2nd International Conference on Agents and Artificial Intelligence - Artificial Intelligence*, pages 527-531
DOI: 10.5220/0002709705270531

2 FIELD ASSOCIATION WORDS

Field Association (FA) words can identify documents. *Short unit association words* are minimum meaningful units which can not be divided without loss of meaning (Atlam, 2002; Atlam 2006). 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 (Aoe, 1989; Breiman 1994; Elmarhomy 2006; Salton 1983; Salton 1988). 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 (Dozawa, 1999) term dictionary, the field tree contains 14 *main (parent) fields*, 18 *middle fields* and 172 *terminal (child) fields*.

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* (*main-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.

2.3 Constructing *FA Words*

2.3.1 Basic Outline

To construct *FA words*, it is first necessary to extract candidate *FA words* from Corpus files classified manually into related fields. Table 1 shows some extracted candidate *FA words* providing information such as: candidates (*A)* are extracted from field (*B)* at a *frequency* (*x times).*

2.3.2 *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 1, "*pass"* is an *FA word* 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* 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* for fields <Election> and <Entrance Examination>,

Table 1: Sample of *FA Words*/ *Declinable Words* and *Association Fields* with Ranks.

FA Words	Field	Rank	Declinable words	Filed	Rank
pass	<Baseball> <Soccer>, <Basketball>	2	*through-pass*	< Soccer >	1
game	<SPORTS>	3	*a perfect game*	<Baseball >	1
Recommendation	<Election>, <Entrance Examination>	4	*recommendation recruitment*	<Entrance Examination>	1
fish	No Association Field	--	*extinction of fish*	<Environment Problem>	1

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. But the condition "*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>.

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 either 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.

3 CONCURRENT WORDS AND ATTACHING WEIGHT

3.1 *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.

3.1.1 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). 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 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 x IDF* is given by:

$$W(d,t) = TF(d,t) \times IDF(t) \quad \text{**********(1)}$$

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

This research applies *TF x 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:

$$\text{Weight}_A(\alpha) = \text{Freq}_A(\alpha) \times \log\left(\frac{N}{\text{Category\_num}(\alpha)}\right) \quad (2)$$

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.

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

$$\text{Weight}_A(\beta) = \text{Freq}_A(\beta) \times \log\left(\frac{N}{\text{Category\_num}(\beta)}\right)$$

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

$$\text{Weight}_A(\alpha+\beta) = \text{Weight}_A(\alpha) + \text{Weight}_A(\beta) = \quad (3)$$
$$\text{Freq}_A(\alpha) \times \log\left(\frac{N}{\text{Category\_num}(\alpha)}\right) + \text{Freq}_A(\beta) \times \log\left(\frac{N}{\text{Category\_num}(\beta)}\right)$$

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} (*foreigner* and *athlete*) =

$$\frac{52 \times \log(133/35) + 535 \times \log(133/57)}{52 \times} = 58.27 \qquad 26$$

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>.

4 EVALUATION RESULTS

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. The accumulating method is to search titles of articles by using keywords exists in field tree system.

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: *C words* are classified manually according to the possibility of field association.
Step 2: Weighting according to *degree of importance* is attached to *C words.*
Step 3: *Precision (P)* and *Recall (R)* are evaluated as

$$\text{Precision (P)} = \frac{\text{Number of } \textit{CFA words} \text{ in the extracted } \textit{C words}}{\text{Total number of } \textit{C words} \text{ automatically extracted}}$$

$$\text{Recall (R)} = \frac{\text{Number of } \textit{CFA words} \text{ in the extracted } \textit{C words}}{\text{Total Number of } \textit{CFA words} \text{ manually extracted}}$$

Tables 4 and 5 show *P* and *R* of *C words* in field <Soccer> are arranged by weighting according to *degree of importance* and *frequency*.

Table 3: Sample *C words* Arranged by the *Degree of.*

C words	***Degree of Importance***	***Frequency***	***CFA word***
1- election area	8703.733	450	Yes
2- election delegate	8193.303	251	Yes
3- vote counting	941.477	123	Yes
4- government election	479.294	28	Yes
5- class being busy	433.257	17	NO
6- seat obtain	284.514	27	Yes
7- proportion turnout	221.666	2	Yes
8- wide support	193.435	16	No
9- prefecture diet time	159.615	6	No
10- seat assure	124.5	12	Yes

Table 4: *Precision & Recall* in Field <Soccer> by Weighting According to *Degree of 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 5: *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

5 CONCLUSIONS

Document fields can be decided efficiently if there are many rank 1 *FA words* and if the *frequency* rate is high, but generally, there are 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 new 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. Future research could focus on clustering *C words* and *FA words*.

REFERENCES

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

Atlam, E.-S., Morita K., and Aoe, J. 2002. A New Method For Selecting English Compound Terms and its Knowledge Representation. *Information P. & Management Journal,* Vol. 38, No. 6, pp. 807-821.

Atlam, E.-S., Morita, K., Fuketa M., and Aoe, J. 2006. Automatic Building of New Field Association Word Candidates Using Search Engine", *Information Processing & Management Journal*, Vol.42, No. 4, pp.951-962.

Breiman, L., Friedman, J.H., Olshen R. A. and Stone C.J. 1994. *Classification and Regression Trees*. Chapman Hall.

Callen, J. P. 1994. Passage and level evidence in document retrieval. *In Proc. of the 17th Annual*

International ACM SIGIR Conference on Research and Development in Information Retrieval, pp.302-310

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

Fuhr, N. 1989. Models for retrieval with probabilistic indexing, *Information Processing and Retrieval*, 25 (1), 55-72.

Elmarhomy, G., Atlam, E.-S., Morita, K., Fuketa, M. and Aoe, J. 2006. Automatic Deletion of Unnecessary Field Association Word Using Morphological Analysis, *International Journal of Computer and Mathematics*, Vol. 83, 3, pp 247-262.

Fukumoto, F., Suzuki, Y. 1996. Automatic Clustering of Articles using Dictionary definitions. Proceeding of the *16th International Conference on Computional Linguistic* (COLING'96), 406-411.

Iwayama, M. and Tokunaga, T., 1999. Probabilistic Passage Categorization and Its Application. *Journal of Natural language Processing*. Vol. 6 No. 3, pp. 181-198.

Melucii, M. 1998. Passage Retrieval and a Probabilistic technique. *Information Processing and Management*, 34(1), 43-68.

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

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

Tsuji, T., Nigazawa, H., Okada, M. and Aoe, J. 1999. Early Field Recognition by Using Field Association Words. In Conference on Computer Processing of Oriental Language, Vol. 2, pp. 301-304.