New Approach for Field Association Term Dictionary with Passage Retrieval

Elsayed Atlam, Elmarhomy Ghada, Masao Fuketa, Kazuhiro Morita and Jun-ichi Aoe
Department of Information Science and Intelligent Systems
University of Tokushima, Tokushima 770-8506, Japan.

Abstract: Field Association (*FA*) *terms* are a limited set of discriminating terms that can specify document fields. Document fields can be decided efficiently if there are many relevant *FA terms* in that documents. An earlier approach built *FA terms* dictionary using a *WWW* search engine, but there were irrelevant selected *FA terms* in that dictionary because that approach extracted *FA terms* from the whole documents. This paper proposes a new approach for extracting *FA terms* using passage (portions of a document text) technique rather than extracting them from the whole documents. This approach extracts *FA terms* more accurately than the earlier approach. The proposed approach is evaluated for 38,372 articles from the large tagged corpus. According to experimental results, it turns out that by using the new approach about 24% more relevant *FA terms* are appending to the earlier *FA term* dictionary and around 32% irrelevant *FA terms* are deleted. Moreover, precision and recall are achieved 98% and 94% respectively using the new approach.

Keywords: Field Association Terms, Passage Retrieval, *WWW* Search Engine, *FA Terms* Dictionary, Recall, Precision.

1- Introduction

In recent years, there has been a tremendous growth of online text information related to digital libraries, medical diagnostic systems, remote education, news sources and electronic commerce. There is a great need to search and organize huge amounts of information in text documents [7][12][16][20][21].

User overload can be reduced and retrieval effectiveness enhanced by retrieving text passages instead of full documents whenever the query similarity of a text excerpt is larger than the query similarity of the complete document [13][16][22].

Field Association (*FA*) *terms* are a limited set of discriminating terms that can specify document fields. For example, term "home-run" can indicate the document field <Baseball>. The basic concept underlying *FA terms* involves choosing a limited set of terms that brings the best matches to a given document. Techniques based on *FA term*s can recognize fields without considering the whole text [10]. Furthermore, document fields can be decided efficiently if there are many relevant *FA terms*.

The main challenge in *FA terms* lie in its extraction and building of a comprehensive *FA term* dictionary. The traditional methods [10] relied on manual extraction of document corpus, but it caused misleading redundant words to be registered because the quality of the resulting *FA terms* depends on the static classified documents by hand. Therefore, Atlam et al. [4] classified documents by using search engine to append *FA terms* candidates to *FA terms* dictionary dynamically. But there were irrelevant selected *FA terms* in that dictionary because that approach extracted *FA terms* from the whole documents. This paper describes improvement of the earlier *FA terms* dictionary by extracting *FA terms* using a passage technique rather than extracting them from the whole documents. The new approach is based on Salton's passage techniques [21] to extract *FA terms* from passages.

2. Field Association Terms

2.1 Field Tree

A field tree is a scheme that represents relationships among document fields.

Figure 1 shows an example of a field tree [6][8], contains 393 terminal fields (<Tennis>, <Golf>, <Fishing>, etc) and 14 medium fields (<Sports>, <Education>, etc.). In Figure 1, the term ("pitching" (231)) in the round brackets of <Baseball> means a *FA term* candidate and its normalized frequency, respectively. The normalized frequency will be discussed in subsection 2.3.

2.2 *FA Term* Scopes

Since *FA terms* have different scope to associate with the field as in Figure 1. Five Scopes are defined to classify *FA terms* to document fields [2][3][4][10] as follows:

Definition 1:
(1) *Perfect-FA terms* are associated with one terminal field (e.g. "pitching" is associated with one terminal field <Baseball>).
(2) *Semi-perfect FA terms* are associated with more than one terminal field in one super-field (e.g. "doubles" is associated with terminal fields <Tennis> and <Table Tennis> of super-field <Sports>).

(3) *Medium-FA terms* are associated with one super-field (e.g. "game" is associated with super-field <Sports>).
(4) *Multiple-FA terms* are associated with more than one terminal field of more than one super-fields (e.g. "victory & defeat" is associated with super-field <Sports> and terminal field <Politics/Election>).
(5) *Non-Specific FA terms* do not specify terminal field or super-fields and also include stop words (e.g. articles, prepositions, pronouns).

2.3 The Determination Algorithm of *FA terms*

This subsection explains the traditional algorithm that automatically determines the candidates for *FA terms* and their scopes [2][3][4][10]. In this algorithm, normalized term frequency is used instead of term frequency in each field as follows:

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

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

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

Definition 2:

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

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

Let *FADIC* be the *FA terms* dictionary. The following algorithm builds the *FADIC* dictionary according to five scopes:

2.4 Drawbacks of the Earlier Approaches

Table 1 shows an example for some extracted terms with their normalized frequencies by using the earlier method.

From Table 1, the earlier method extracted *FA terms* such as "*season*" and "*veteran*", "*lukewarm*", and "*trading*" from the whole documents. These *FA terms* are appended to the existing dictionary as they have high normalized frequencies compared with the total number of <Baseball>, but these *FA terms* were not extracted by the earlier approach because they have low frequency compared with the total number of extracted *FA terms* in the whole text. extracted *FA terms* in the whole text. But The main challenge in *FA terms* lie in its extraction and building of a comprehensive *FA terms* dictionary. The traditional methods [9][23] relied on manual extraction of document corpus, but it caused misleading redundant words to be registered because the quality of the resulting *FA terms* depends on the static classified documents by hand. Therefore, Atlam et al. [4] classified documents by using search engine to append *FA term* candidates to *FA terms* dictionary. Atlam's method extracts *FA terms* automatically as follows:

(*Step a*) Normalized term frequencies are determined for the whole documents.

(Step b) FA term scopes are determined by using normalized frequency information.

(Step c) FA terms are automatically appended in an existing *FA terms* dictionary.

normalized term frequencies are determined for *FA term* candidates ("*season*", (223)), ("*contracts",* (22)), ("McGriff", (240)),("*trading*", (242)), ("lukewarm", (235)), ("baseball", (262)), ("pitching", (64)), ("lineup", (69)), ("baseman:, (238)), ("*veteran*", (227)), and ("home runs", (81)) as shown in Table 1. In (*Step b*), *FA term* scopes are determined by using Concentration ratio and static approach. These *FA terms* are appended dynamically to the existing *FA terms* dictionary in (*Step c*).

Table 1: Sample normalized term frequencies using Atlam's method

Terms	Norm. term freq. Terl. Fel.	Norm. term freq. Med. Fel	Extracted (✓) /not extracted (×)	relevant (*r*) / irrelevant (*i*)
season	**223**	**296**	✓	***i***
contract	**22**	**327**	×	***i***
McGriff	**240**	**254**	✓	***r***
trading	**242**	**267**	✓	***i***
lukewarm	**235**	**242**	✓	***i***
baseball	**262**	**275**	✓	***r***
pitching	**64**	**194**	×	***r***
lineup	**69**	**189**	×	***r***
baseman	**238**	**243**	✓	***r***
veteran	**227**	**230**	✓	***i***
home runs	**81**	**245**	×	***r***

these *FA terms* are not restricted to the specific field <Baseball>.

Moreover, *FA terms* "*pitching*" and "*lineup*" are relevant and restricted to the field

Figure 2 shows an example of document paragraphs denoted by *p* and *p'* etc. In order to explain the earlier method [5].

Thus the earlier approach has two drawbacks:

(I) Irrelevant selection of *FA terms* which are not restricted to the specific field.

(II) Non-selection of *FA terms* which are restricted to the specific field.

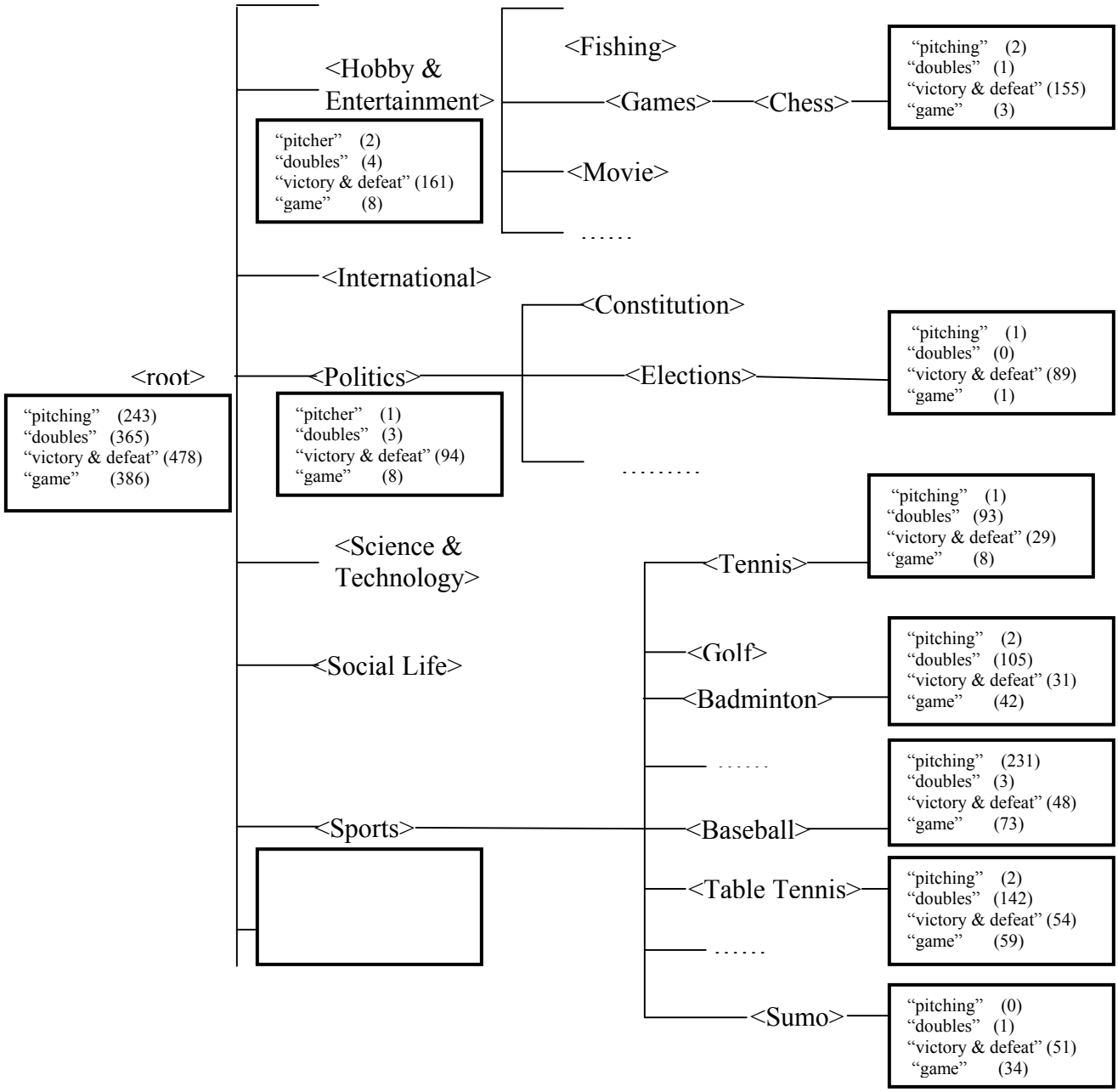

Fig. 1: An example of Field tree and *FA terms*

Therefore, the next section describes a new approach to resolve these disadvantages of earlier method by using passage retrieval.

3. Improvement of *FA Terms* Extraction by using Passage Retrieval

3.1 Passage Retrieval Technique

This section introduces Salton's passage retrieval technique (Salton et al., 93)

Salton's passage retrieval technique is capable of retrieving relevant texts with a high degree of accuracy in response of a user's query. This technique extracts text paragraphs that are relevant to the user queries. Let $Sim(d_i d_j) = \sum_{k-1}^{t} w_{ik} w_{jk}$ be the *stander inner product vector* similarity function which produces similarity coefficient between 0 and 1 that depend on the proportion and the weight of matching terms w_{ik} and w_{jk} in the two document texts d_i and d_j.

Analysis of documents

"Vaughn joining Canseco"

…….

[paragraph p]

Devil Rays **adding *pop* at the *Trop*. The Tampa Bay Devil Rays lost 1 million fans last** season **trying to *win* with "*pitching*" and** defense**. Now, with** Vinny Castilla **and Greg Vaughn joining Jose Canseco and Fred McGriff in their "*lineup*", they are at least trying to be interesting, if not *competitive*.** Castilla, Vaughn, McGriff **and** Canseco **hit 144 "*home runs*", last year, but they're not exactly known for their youth, or their *glove work*. *Castilla,* who turns 33 next year, is the youngest of them. The Rockies were concerned that *Castilla's bat,* and his range at third slipped noticeably last *season*. The** Devil Rays **have long-term money sunk into this radical change in philosophy. With a need to put baseball on the map in the** Tampa Bay **area, they are hoping fans like the idea of watching** veteran **sluggers take aim at the Tropicana Field catwalks.**

…….....

[paragraph p']

Jay not walking Toronto **signs 1B Delgado to 3-year contract. The power-hitting first** baseman **agreed** Friday **to a $36 million, three-year contract the Toronto Blue Jays. But the team gave him the right to demand a trade after next** season. **"We were able to work a deal out where I was happy, not so much with the money but with some of the clauses that are in the** contract**,"** Delgado **said. "At the end of** season**, we were talking about a long-term contract but I had a few concerns." "I'm not going to jump *the gun* and say, well in September things are going to be bad. And I don't see a reason why things are going to be bad with Fregosi." The Blue Jays discussed** trading **Delgado after *contracts* talks stalled, but demand for Delgado on** baseball's trade market **was *lukewarm*, with four teams believed to have expressed interest. Clubs were reluctant to trade quality players to Toronto and then sign Delgado to a huge contract. "This is** pretty **awesome from *the business* side,"** Delgado **said. "I love the city, the money is right, the terms are right and the length is right. In three years I'm a free agent again.**

Fig. 2: Document analysis by the earlier method

In this paper, Salton's method is defined as the following function SALTON(*d*) that produces a set of passages for the input query document *d*.

The function SALTON(*d*):

Figure 4 shows an example of Slaton's approach and how to replace the whole document text on the output list by the corresponding text paragraphs. The text of a typical query article ([195], Vaughn) is shown in Figure 4 (a).

This article consists of a single text paragraph. A typical article retrieved among the top 100 is [1350] entitled "Vaughn joining Canseco". The article consists of three paragraphs (labeled [1350.p], [1350.p'] etc.. The text and vector similarities as well as the maximum sentence similarities between query and document paragraphs are shown in Figure 4 (b). In this case, two paragraphs have a higher global query similarity than the full documents (paragraph [1350.p] and [1350.p']). However, the second of these does not contain any sentence with the required sentence similarity with the query. Hence, paragraphs p' is discarded, and paragraph p is used for the retrieval purposes in this case. The text paragraph [1350.p], entitled, "Vaughn and McGriff hit 144 homers last year" is shown in Figure 4 (c). Figure 4 shows that the retrieved text paragraph [1350.p] is correspondingly closely related to the query article [195].

3.3 Examples of New approach using Salton's Technique

The new approach in this paper combines Atlam's and Salton's methods by applying Atlam's approach on the passages determined by Salton's instead of using whole documents. For example, Figure 2 shows a sample of extracted passage *p* from whole documents after using Salton's technique. In the other side some irrelevant text excerpts p_i are deleted with many irrelevant terms "*trade market*", "*lukewarm*", "*pretty*" etc that were extracted from the whole documents. Moreover, after applying Atlam's method on the extracted passage p_i the normalized term frequencies of ("*pitching*", (43), ("*lineup*", (39)), ("*home runs*", (44)) becomes high compared with the total frequencies and these terms extracted as *FA terms* shown in Table 2. Therefore, *FA terms* "*pitching*", "*lineup*", "*home runs*" are extracted by

using the presented method, but those were not considered by the earlier approach.

Table 2 extracted *FA terms* using new approach

Term	Nor. term freq. <*Terminal Field*>	Nor. term freq. in <Medium Field>	extracted (✓) / not extracted (✖)	relevant (*r*) / irrelevant (*i*)
season	**7**	**23**	✓→✖	***i***
McGriff	**46**	**50**	✓	***r***
baseball	**45**	**49**	✓	***r***
pitching	**43**	**47**	✖→✓	***r***
lineup	**39**	**42**	✖→✓	***r***
veteran	**6**	**20**	✓→✖	***i***
Home runs	**44**	**48**	✖→✓	***r***

✖→✓ = means some un-extracted terms is extracted by the n.
✓→✖ = means some extracted terms is un-extracted by the

Thus, the new approach can append more relevant *FA terms* to the existence dictionary built by the earlier method.

Table 2 shows an example for some extracted *FA terms* with their frequencies by using passage technique. From Table 2 we notice that some relevant *FA terms* are extracted by using the new technique which was not extracted by the earlier approach as they have low frequency.

3.4 The Presented Algorithm Using Passage Techniques

A new algorithm of appending FA terms automatically can be summarized as follows:

Input: (a) A field <*F*>
(b) A set of reference keywords REFKEY(<*F*>), *perfect FA terms,* in a given field <*F*> from the *FA term FADIC* dictionary.
(c) Document passages

Output: The extended *FA term* dictionary *FADIC.*

Method:

(Step 1): Determine a set of reference keywords REFKEY(<*F*>) (perfect *FA* terms) from the dictionary *FADIC* for a given field <*F*>.

(Step 2): For REFKEY(<*F*>), determine a set *D* of retrieved documents by using a *WWW* search Engine.

(*Step 3):* For each document *d* in *D*, determine a set *P* of relevant passages from the function SALTON(*d*).

(Step 4): All relevant passages *p* in *P* are appended into the classified documents for <*F*>.

(Step 5): Determine the extended dictionary *FADIC* by calling the algorithm *FABUILD* for the field tree with the new documents for <*F*> in (Step 4):

4. The Experimental Evaluation

4.1 The Experimental Data

In the experimental evaluation, about 14.52MB document data have been collected by using Google search engine. This collection corpus contains 16,400 text files of 850 classified fields related to various topics as <Sports>, <Computers>, <Politics>, <Economics>, etc. From this collection, around 38,372 term candidates from fields are used for the experimental evaluation. Concentration ratio is 0.5 ~ 0.9 to determine *FA term* candidates.

4.2 The Accuracy Evaluation

Precision and *Recall* used to evaluate the new approach (Salton and McGill, 1983) is defined:
Suppose that x = Number of relevant *FA terms* extracted automatically by using the new approach y = Total Number of *FA terms* extracted automatically by using the new approach z = Total Number of relevant *FA terms* extracted manually

Therefore $Precision\ (P) = \frac{x}{y}$, $Recall\ (R) = \frac{x}{z}$

4.3 Experimental Simulation Results

This section shows the effectiveness of using document passages for extracting *FA terms*. Figure 3 shows that the number of extracted *FA terms* increases using new approach leading to increase in relevant *FA terms* related to the field. Moreover, the number of extracted *FA terms* increases as concentration ratios decrease. When the concentration ratio is high, *FA terms* are extracted accurately, but when the ratio is low, *FA terms* are not accurately determined. From Fig. 3, it is clear that the number of relevant *FA terms* increases as the number of extracted *FA terms* increases too. Moreover, *FA terms* becomes accurate with the increases of the concentration ratio therefore concentration ratio 0.9 is the most effective.

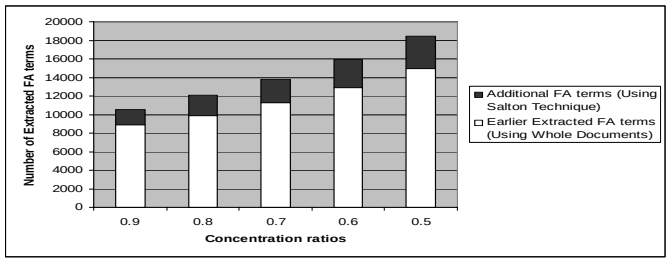

Fig. 3 Concentration ratios with *FA terms* and additional *FA terms* after using new approach

5. Conclusion

This paper has described improvement of an earlier *FA terms* dictionary by extracting *FA terms* using Salton's passage technique rather than extracting them from the whole documents. The advantage of the new approach is discriminating more relevant *FA terms* as the passage retrieval techniques substantially furnish output better than the whole documents. The proposed approach has been evaluated for 38,372 articles from the large tagged corpus. According to experimental results, it turns out that by using the new approach about 24% more relevant *FA terms* are appending to the earlier *FA term* dictionary and around 32% irrelevant *FA terms* are deleted. Moreover, *precision* and *recall* are achieves 98% and 94% respectively by using new approach.

Future work could focus on document summarization based on the approach obtained from the proposed method and *FA terms* information could be applied.

References

[1] J. Aoe, K. Morita & H. Mochizuki, H. An Efficient Retrieval Algorithm of Collocate Information Using Tree Structure. Transaction of the IPSJ, Vol. 39, No. 9, 1989, pp. 2563-2571.

[2] E.-S. Atlam, K. Morita, M. Fuketa & J. Aoe. A New for Selecting English Compound Terms and its Knowledge Representation. Information Processing & Management Journal, Vol. 38, No. 6, 2002, pp. 807-821.

[3] E.-S. Atlam, M. Fuketa, K. Morita & J. Aoe. Document Similarity measurement using Field association terms. Information Processing & Management Journal, Vol. 39, No. 6, 2003, pp. 809-824.

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

[5] M. Bacchin, N. Ferro and M. Melucci. A probabilistic model for stemmer generation, *Information Processing & Management*, Vol. 41, No. 1, 2005, pp.121-137.

[6] A. Bratko and B. Filipič. Exploiting Structural Information for Semi-Structured Document Categorization, *Journal of Information Processing Management, Vol.* 42, No. 3, 2006, pp. 679-694.

[7J.P. Callan. Passage and Level Evidence in Document Retrieval. In Proc. of the 17th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, 1994, pp. 302-310.

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

[9] M. Dumais, E. Banko, J. Brill and A. Linm. Web Question Answering: Is More Always Better. *In Proceedings of ACM SIGIR 2002*, pp. 291-298.

[10] M. Fuketa, S. Lee, T. Tsuji, M. Okada and J. Aoe. A Document Classification by using Field Association Words. *International Journal of Information Sciences*, Vol.126, 2000, pp. 57-70.

[11] F. Fukumoto, Y. Suzuki. Automatic Clustering of Articles using Dictionary definitions. In roceeding of the 16th International Conference on Computational Linguistic (COLING'96), 1996, pp. 406-411.

[12] M.A. Hearst, & C. Plaunt. Subtopic structuring for full-length document access. In HARD 2004-Passage Retrieval Using HMMs, University of Illinois at Urbana-Champaign, TREC 2004, 1993.

[13] M. Kaszkiel & J. Zobel. Passage retrieval revised In Proc. of the 20th Annual International ACM SIGIR Conference on Research and Development in information Retrieval, 1997, pp.178-185.

[14] R. Korfhage, E. Rasmussen, & P. Willet (Eds.), *Proceedings of the 16th annual international ACM-SIGIR conference on research and development in information retrieval*, Pittsburgh, PA New York: ACM, 59-68.

[15] H. Kyoung, S. Young-In, K. Sang-Bum and R. Hae-Chang. Answer extraction and ranking strategies for definitional question answering using linguistic features and definition terminology, Journal of Information Processing & Management, Vol. 43, No. 2, 2006, pp.353-364.

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

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

[18] S. R. Safavian and D. Landgrebe. A Survey of Decision Tree Classifier Methodology. *IEEE Trans. On systems, Man, and Cybernetics*, Vol. 21, No.3, 1991, pp.660-674.

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

[20] G. Salton, J. Allan and A. K. Singhal. Automatic Text Decomposition and Structuring. *Information Processing and Management*, Vol. 32 No. 2, 1996, pp.127-138.

[21] G. Salton, J. Allan and C. Buckley. 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, 1993, pp. 49-58.

[22] L. Sangkon, M. Shishibori, T. Sumitomo and J. Aoe. Extraction of Field-Coherent Passages. *Journal of Information Processing & Management*, Vol. 38, No. 2, 2002, pp. 173-207.