Computing and Informatics, Vol. 32, 2013, 371–391

AN EFFICIENT METHOD OF SUMMARIZING DOCUMENTS USING IMPRESSION MEASUREMENTS

Abdunabi Ubul, El-Sayed Atlam, Hiroya Kitagawa
Masao Fuketa, Kazuhiro Morita, Jun-ichi Aoe

Department of Information Science and Intelligent Systems
Faculty of Engineering, University of Tokushima
Minami josanjima 2-1
770-8506 Tokushima, Japan
e-mail: {abdunabi211, atlam, kitagawa, fuketa, kam, aoe}@is.tokushima-u.ac.jp

Communicated by Prabhat Kumar Mahanti

Abstract. Automatic generic document summarization based on unsupervised schemes is a very useful approach because it does not require training data. Although techniques using latent semantic analysis (LSA) and non-negative matrix factorization (NMF) have been applied to determine topics of documents, there are no researches on reduction of matrix and speeding up of computation of the NMF method. In order to achieve this scheme, this paper utilizes the generic impressive expressions from newspapers to extract important sentences as summary. Therefore, it has no stemming processes and no filtering of stop words. Generally, novels are typical documents providing sentimental impression for readers. However, newspapers deliver different impressions for new knowledge because they inform readers about current events, informative articles and diverse features. The proposed method introduces impressive expressions for newspapers and their measurements are applied to the NMF method. From 100 KB text data of experimental results by the proposed method, it turns out that the matrix size reduces by 80 % and the computation of the NMF method becomes 7 times faster than with the original method, without degrading the relevancy of extracted sentences.

Keywords: Impressive expressions, NMF methods, precision, relevancy

1 INTRODUCTION

The rapid advancement of Internet has led to a deluge of information on the Web and has caused difficulties to locate required information efficiently. With increasing the availability of information as well as with the fact that there is not enough time to read them, document summarization technologies [1, 2, 3, 4, 5, 6, 7, 8, 9, 10] have become an extremely important area of research. Document summarization can be used in many applications such as information retrieval, intelligence gathering, information extraction, text mining, document similarity text and news broadcasting [11, 12, 13, 14, 15, 16, 17, 18].

Automatic generic document summarization based on unsupervised schemes is a very useful approach because it does not require training data. Zha [19] and Yeh et al. [20] have proposed summarization methods using Latent Semantic Analysis (LSA), but the LSA methods can not extract meaningful sentences because of many features with positive and negative values. Li et al. [21] have solved this problem by introducing generic multi-document summarization and Zha [19] has proposed the mutual reinforcement principle (MRP) to query-based document summarization. However, problems related to extraction of subtopics in summarization have not been solved yet. In order to solve this problem, Hong Lee [22] has proposed a new unsupervised generic document summarization method using a non-negative matrix factorization (NMF) method called the original method in this paper. Semantic feature vectors extracted from NMF can be interpreted more intuitively than those extracted from LSA-related methods. Improvement is estimated by experimental results for DUC 2006 test data [23]. Although these techniques have been applied to determine topics of newspapers, there is no research on reduction of matrix and speeding up of computation for the NMF method.

In order to achieve this scheme, this paper utilizes the generic knowledge of expressions from newspapers to extract important sentences as the summary. Therefore, it has no stemming processes and no filtering of stop words. Generally, novels are typical documents providing sentimental impression for readers. However, newspapers deliver different impressions for new knowledge because they inform readers about current events, informative articles and diverse features. Moreover, newspapers also include editorial pages and columns expressing personal opinions of writers. Although newspapers have many kinds of fields (politics, education, economy, business, entertainment, society, sports, etc.), impact is given based on new facts in common, for all fields. Therefore, the proposed method defines these common expressions for newspapers and builds a generic impressive dictionary using these expressions.[1, 2, 3]

The impression measurements are defined for each item of the impressive dictionary and the measurement is applied to the NMF method. For documents test data

[1] http://eqi.org/fw.htm

[2] http://www.eqi.org/fw_neg.htm

[3] http://www.winspiration.co.uk/positive.htm

from DUC2006 [23] data set, the space, time and quality of summarization is estimated by comparing extracted sentences by the proposed method with those of the original NMF method. In the comparison, the ROUGE software is utilized. From 100 KB text data of experimental results by the proposed method, it turns out that the matrix size is reduced by 80 % and computation of the NMF method becomes 7 times faster than with the original method, without degrading the relevancy of extracted sentences.

Section 2 of this paper describes the outline of the proposed method. Section 3 describes drawbacks of the traditional methods using stemming and stop words, and proposes definitions and measurements for impression expressions. Section 4 describes summarization by the NMF method using impression measurements. Section 5 evaluates the proposed generic document summarization method by comparing with the original NMF method. Section 6 concludes the proposed method and discusses future works.

2 OUTLINE OF THE PROPOSED METHOD

Figure 1 shows the flow of a document summarization system.

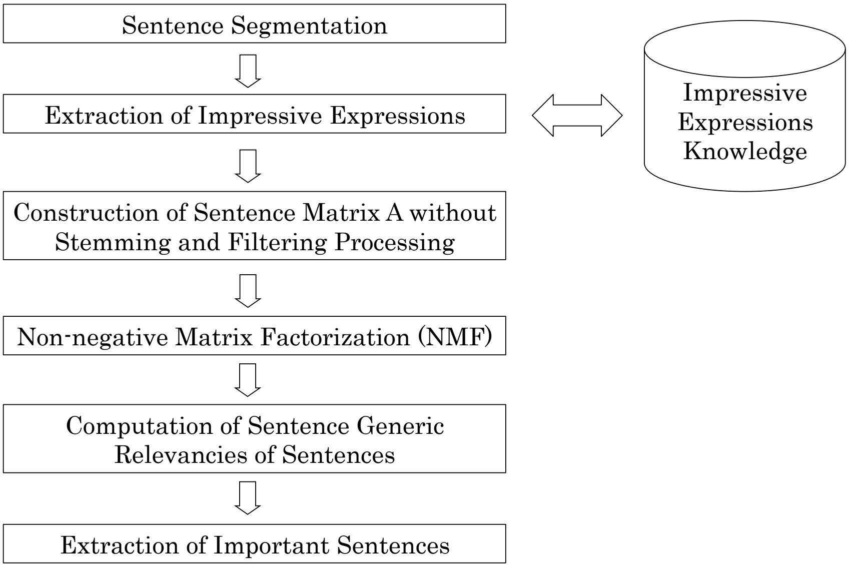

Sentence Segmentation
Extraction of Impressive Expressions
Impressive Expressions Knowledge
Construction of Sentence Matrix A without Stemming and Filtering Processing
Non-negative Matrix Factorization (NMF)
Computation of Sentence Generic Relevancies of Sentences
Extraction of Important Sentences

Fig. 1. Flow of a document summarization

In sentence segmentation modules of Figure 1, all documents are separated into collection of sentences or paragraphs by segmentation modules according to

Hao's [24] approach. Second, impressive expressions are extracted by using impressive expressions dictionary manual. The original NMF method constructs a sentence matrix A using all resulting words of the segmentation module, but the proposed method introduces the extraction module of determining impressive expressions without the stemming process and filtering of stop words. Sentence matrix A is constructed by using the resulting expressions of the above extraction module. Finally, generic relevancies of sentences are computed and important sentences are extracted as the summary.

The impression measurements will be discussed in Section 3 and the NMF module will be discussed in Section 4.

In this paper, the proposed method uses the same document as shown in Table 1. Table 1 shows segmented sentences for document D0601A from corpus DUC2006 [23]. Document D0601A is related to the topic *"Native American Reservation System – pros and cons"*. In Table 1, the segmented sentences are represented by $S_1, S_2, \dots, S_3$. Underlined words are representing impressive expressions which will be explained in the following sections.

S_i	Sentences
S_1	President Clinton turned the attention of his national poverty tour today to arguably the poorest, most forgotten U.S. citizens of them all: American Indians.
S_2	Clinton was going to the Pine Ridge Reservation for a visit with the Oglala Sioux nation and to participate in a conference on Native American homeownership and economic development. He also was touring a housing facility and signing a pact with Oglala leaders establishing an empowerment (Fixing his position) zone for Pine Ridge.
S_3	But the main purpose of the visit – the first to a reservation by a president since Franklin Roosevelt – was simply to pay attention to American Indians, who are so raked by grinding poverty that Clinton's own advisers suggested he come up with special proposals geared specifically to the Indians' plight.
S_4	At Pine Ridge, a scrolling marquee at Big Bat's Texaco expressed both joy over Clinton's visit and wariness of all the official attention: "Welcome President Clinton. Remember Our Treaties," the sign read.
S_5	According to statistics from the Census Bureau and the Bureau of Indian Affairs, there are 1.43 million Indians living on or near reservations. Roughly 33 percent of them are children younger than 15, and 38 percent of Indian children aged 6 to 11 live in poverty, compared with 18 percent for U.S. children of all other races combined.
S_6	Aside from that, only 63 percent of Indians are high school graduates. Twenty-nine percent are homeless, and 59 percent live in substandard housing. Twenty percent of Indian households on reservations do not have full access to plumbing, and the majority – 53.4 percent ? do not have telephones.
S_7	The per capita income for Indians is $ 21 619, one-third less than the national per capita income of $ 35 225. An estimated 50 percent of American Indians are unemployed, and at Pine Ridge the problem is even more chronic – 73 percent of the people do not have jobs.
S_8	Housing Secretary Andrew Cuomo, who visited the reservation last August, said Pine Ridge is a metaphor for the poverty tour, for it sits in Shannon County, the poorest census tract in the nation.
S_8	This is generations of poverty on the Pine Ridge reservation, with very, very little progress, Cuomo said. "We didn't get into this situation in a couple of weeks and we're not going to get out of it in a couple of weeks. It's going to take years."
S_{10}	To begin addressing the housing problem, Clinton was announcing a partnership between the Treasury Department, the Department of Housing and Urban Development, tribal governments and mortgages companies to help 1 000 Indians become homeowners over the next three years – a small number that none the less would double the number of government-insured home mortgages issued on tribal lands. Under the effort, "one-stop mortgage centers" would be opened at Pine Ridge and on the Navajo Reservation in Arizona to help streamline the mortgage lending process.
.	....
S_{14}	The announcement was part of Clinton's four-day, cross-country tour to highlight the "untapped markets" in America's inner cities and rural areas.

Table 1. Examples of segmented sentences for documents from DUC2006

3 DRAWBACKS AND MEASUREMENTS OF IMPRESSIONS

3.1 Original Method Drawbacks

a) Stemming Drawback

In traditional methods, one of the major defects of stemming like Porter's [25] is that they often conflate words with similar syntax but completely different semantics. For example, "news" and "new" are both stemmed to "new" while they belong to two quite different categories. Moreover, words "USA" and "ADIS" become "U" and "AD", respectively after stemming, which have no meaning or relation to the original one at all in the sentences. This paper proposes the generic knowledge of expressions for solving this problem.

b) Stop List Drawback

Disadvantage of the original method is that a defined list of stop words consists of many words having significant meanings in texts which helps extract a correct summary. For example, contrastive conjunction words "however", "because", and "but" are very important to derive conclusions in context and they support to evoke reader's impression indirectly. Therefore, in this paper, we use a summarization process without filtering stop words.

Table 2 shows results of stemming and stop words results from the segmented sentences of document D0601A from DUC corpus.

From Table 2, it is clear that, after stemming, some words become conflated words with similar syntax but completely different semantics than original ones such as word "u" in sentence S1. Moreover, some stop words have significant meanings in texts such as "less" and "up" which represent objective comparisons to support reader's impression implicitly.

Therefore, this study has no stemming processes and no filtering of stop words; but the size of matrix A becomes large by using all words. In order to reduce the size of matrix A and to speed up computation of the NMF method, the generic of impressive expressions will be proposed in the following sub-sections.

3.2 Measurements of Impressions

3.2.1 Definitions of Impressive Expressions

There are many methods to extract popularity and non-popularity for CGM managements based on reputation including emotional expressions [26, 27, 28]. Although the aim of these researches is not to summarize newspapers, the concept by positive and negative expressions is also used for impressive knowledge to be proposed here. This paper defines impressive expressions as follows:

Sentences	Stemming word	Stop word
S_1	arguabl, attent, citizen, hi, Indian, nation, poverty, presid, ridg, today, turn, u	all, his, most, of, the, them, to
S_2	confer, develop, econom, empower, establish, facil, go, hous, leader, nativ, particip, reserv, ridg, sign, tour	He, a, also, an, and, for, in, on, the, to, was, with
S_3	advis, attent, clinton, gear, grind, indian, pai, poverti, presid, propos, purpos, rake, reserv, simpli, sinc, specif, suggest	But, a, are, by, first, he, of, own, since, so, that, the, to, up, was, who, with
S_4	Attent, bat, Clinton, express, joi,marque, offici, presid rememb, ridg, scroll, treati, wari, welcom	At, Our, a, all, and, at, both, of, over, the
S_5	accord, affair, ag, censu, combin, compar, indian, live, poverti, race, reserv, roughli, statist, u	all, and, are, for, from, in, of, on, or, other, than, the, them, there, to, with
S_6	asid, graduat, hous, household, indian, major, plumb, reserv, telephon, twenti, twenti-nin	and, are, from, have, in, not, of, on, only, that, the, to
S_7	estim, incom, indian, job, nation, on-third, peopl, ridg, unemploi	An, The, and, are, at, even, for, have, is, less, more, not, of, per, than, the
S_8	censu, counti, hous, poverti, reserv, ridg, secretari, sit, visit	a, for, in, is, it, last, the, who
S_9	coupl, gener, go, it, littl, poverti, reserv, ridg, situat, veri, week, year	It's, This, We, a, and, in, into, is, it, not, of, on, out, the, this, to, very, we're, with
S_{10}	address, announc, becom, center, compani, depart, develop, doubl, govern, govern-insur, homeown, hous, indian, issu, land, lend, mortgag, open, reserv, ridg, streamlin, treasuri, year	To, Under, a, and, at, be, become, between, in, next, of, on, over, that, the, to, was, would
.	....	....
S_{14}	announc, clinton, four-dai, cross-countri, untap, market, america, citi, area	and, of, to, in, The, was, the

Table 2. Results of stemming and stop words from the segmented sentences in Table 1

A) Explicit Expressions

a1) Emotional Expressions

Although emotional expressions represent subjective personal impressions in general, the expressions in newspapers have objective impressions. For example, it is intuitively clear that "happy" has positive (p) impression and "sad" has negative (n) impression. Of course, it is difficult to classify some of them into positive and negative impressions when the determination depends on the context. It is defined as (pn) impression. For example, "surprise" has (pn) impression because it can be used both in positive and negative situations.

a2) Strong Associative Expressions

These expressions can associate with positive and negative impressions strongly. For example, "peace" and "healthy" affect readers by positive impressions. On the other hand, "war" and "disease" affect humans by negative impressions. For example, in Table 1, S_1 *"President Clinton turned the* ***attention*** *of his national* ***poverty*** *tour today to arguably the* ***poorest****, most* ***forgotten*** *U.S. citizens... etc."*), the impressive term ***"attention"*** refers to strong positive (ps) while ***"poorest"*** and ***"forgotten"*** terms refer to strong negative (ns).

a3) Weak Associative Expressions

These expressions can associate with positive and negative impression weakly. For example, "buy" and "full" have positive impressions [28], and "weak" and "drop" have negative impressions [26, 27], but the power of association is weak. There are many candidates and one of them depends on positive actions "visit, reply, establish" and negative actions "stop", "get out".

B) Implicit Expressions

b1) Evaluative Expressions

These expressions indirectly derive positive and negative impressions.

b11) Comparative Expressions

These expressions represent objective comparisons to support reader's impression implicitly. Examples are "more", "less", "than", "comparison" and "rate", as in Table 1, S_5 *"Roughly 33 percent of them are children younger* ***than*** *15".* Although "good", "poor" and "bad" also belong to this class, they are defined as the above explicit class.

b12) Numerical Expressions

These expressions represent numerical results to support reader's impression implicitly. For example, "percent", "as-as", "rate", and "result", as in Table 1, S_7 *"73* ***percent*** *of the people do not have jobs."*

b2) Conjunctional Expressions

These expressions associate with special reasons, explanations and conclusions in context and they support evoking reader's impression indirectly. Specially, contrastive conjunctions "however", "but" and "on the other hand" are very important to derive conclusions, for example in the sentence *"a member of the House leadership and a smoker himself, said the bill would seek to reduce underage smoking,"* ***but*** *he added: "Teen-agers are going to smoke."* Another example is shown in S_3, Table 1.

b3) Adverbial Expressions

These expressions enhance sentences to make the news items impressive. Emphatic adverbs "specially" and "strongly", are important. These expressions do not exist in Table 1, but they are included in another document sentences.

In this paper, the following abbreviations will be used:

- *ps, pw* representing strong and weak positive impressive expressions,
- *ns, nw* representing strong and weak negative impressive expressions,
- *pnw* can associate with positive and negative impressive expressions weakly,
- *es, ew* can associate with strong and weak evaluation expressions.

4 SUMMARIZATION ALGORITHM

This section proposes a method to create generic document summaries by selecting sentences using NMF. The proposed method consists of a preprocessing step and a summarization step as in the following sub-sections.

4.1 Non-Negative Factorization (NMF)

NMF is introduced by basic notations [30, 31]. NMF decomposes word-sentence matrix $A(n \times m)$ size into two matrices of $W(n \times k)$ size and $H(k \times m)$ size for k such that $k < n$ and $k < m$ as follows:

$$A \approx WH \tag{1}$$

where W is called a non-negative semantic feature matrix (NSFM) and H is called a non-negative semantic variable matrix (NSVM). Let $A[i,j]$, $W[i,j]$ and $H[i,j]$ be elements of A, W and H for the i^{th} row and for the j^{th} column, respectively. $W[*,j]$ and $H[*,j]$ represents a semantic feature and semantic variable vectors for the j^{th} column, respectively.

In order to satisfy the approximation condition $\tilde{A} = WH$, the following Frobenius norm is used [32, 33]:

$$\Theta_E(W,H) \equiv \|A - WH\|_F^2 \equiv \sum_{i=1}^{m}\sum_{j=1}^{n}\left(A[i,j] - \sum_{h=1}^{k} W[i,h]H[h,j]\right)^2 \tag{2}$$

$\Theta_E(W,H)$ is computed until it exceeds the number of repetitions or the predefined threshold, using the following updating rules:

$$H[i,j] \leftarrow H[i,j]\frac{(W^T)A[i,j]}{(W^TWH)[i,j]}, \quad W[i,j] \leftarrow W[i,j]\frac{(AH^T)[i,j]}{(WHH^T)[i,j]} \tag{3}$$

where W^T, H^T, representing the transpose of matrices W and H, respectively, formed by turning rows into columns and vice versa. By the recomposing process, the sentence vector $A[*,j]$ can be represented by a linear combination of the h^{th} semantic feature vectors $W[*,h]$ and the semantic variable $H[h,j]$ as follows:

$$A[*,j] = \sum_{h=1}^{k} H[h,j]W[*,h] \tag{4}$$

Table 3 shows a sentence matrix A for words and sentences obtained by the preprocessing applied on a set of sentences in Table 1, where FW means impressive expressions and IM means impression measurement. In Table 1, Matrix A is corresponding to 79 terms of 14 sentences.

N_0	FW	IM	S_1	S_2	S_3	S_4	S_5	S_6	S_7	S_8	S_9	S_{10}	$\cdots$	S_{14}
1	attention	ps	1	0	1	0	0	0	0	0	0	0	...	0
2	country	pw	0	0	0	0	0	0	0	1	0	0	...	0
3	majority	ps	0	0	0	0	0	1	0	0	0	0	...	0
4	problem	ns	0	0	0	0	0	0	1	0	0	1	...	0
5	percent	es	0	0	0	0	3	5	2	0	0	0	...	0
6	leader	ps	0	1	0	0	0	0	0	0	0	0	...	0
7	poverty	ns	1	0	1	0	0	0	0	1	1	0	...	0
8	poorest	ns	1	0	0	0	0	0	0	1	0	0	...	0
9	job	pw	0	0	0	0	0	0	1	0	0	0	...	0
10	suggest	ps	0	0	1	0	0	0	0	0	0	0	...	0
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
79	help	ps	0	0	0	0	0	0	0	0	0	2	$\cdots$	0

Table 3. Sentences matrix A for document from DUC2006

In Table 3, matrix A, the rows represent 79 of impressive expressions and the columns represent 10 segmented sentences, while the i^{th} rows and the j^{th} columns represent term frequency (i.e.... $A[6,5] = 3$).

Table 4 explains the semantic feature of sentences applying NMF to matrix A, where W_i means $W[*, i]$. The original method is using 396 terms of 57 sentences while the proposed methods is using 79 terms of 14 sentences which reduces the matrix A size and speeds up the processing in the following section.

N_0	FW	Semantic feature											
		W_1	W_2	W_3	W_4	W_5	W_6	W_7	W_8	W_9	W_{10}	$\cdots$	W_{14}
1	attention	0	0	0	0	13.32	0	0	0	0	0.01	$\cdots$	19.79
2	country	0	0	8.12	6.59	0	0	0	0	0	0	$\cdots$	0
3	majority	3.86	0	0	0	0	0	0	0	0	0	$\cdots$	0
4	problem	0	4.95	0	0	0.04	6.35	4.64	0	0.29	0.02	$\cdots$	0
5	percent	0	4.89	0	0	0	0	0	0	0	0	$\cdots$	0
6	leader	0	4.89	0	0	0	0	0	4.59	0	0	$\cdots$	0
7	poverty	0	0	0	0	0	5.32	0	0	0	0	$\cdots$	0
8	poorest	0	0	4.11	0.02	0	0	0	0	0	0	$\cdots$	0
9	job	0	0	0	0	0	5.32	0	0	0	0	$\cdots$	0
10	suggest	0	0	8.23	0.02	0	0	0	0	0	0	$\cdots$	0
$\cdots$	$\cdots$.$\cdot$.$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$
79	help	0	0	0	0	0	0	9.07	0	4.66	0	$\cdots$	0

Table 4. Semantic feature of sentences using NMF

In Table 4, the semantic feature vectors, $W_1, W_2, \ldots, W_{14}$, are obtained from NMF decomposition of matrix A are shown. From Table 3, it is clear that the highest semantic feature values are $W[3,3]$, $W[3,4]$, $W[5,6]$, and $W[8,6]$ with the impressive expressions "country", "problem", and "poverty" respectively, which will be affected in extracting important sentences as the summary.

Table 5 shows the semantic variable vectors for sentences obtained by NMF to matrix A, where H_i means $H[*, i]$.

In Table 5, the semantic variable vectors $H_1, H_2, \ldots, H_{10}$ obtained from NMF decomposition of matrix A are shown. From Table 4, it is clear that the semantic variable vector $H[10,1]$ is representing the highest value. This value will be used in Generic Relevance of a Sentence (GRS) for document summarization in the following section.

Semantic variable	Sentence											
	S_1	S_2	S_3	S_4	S_5	S_6	S_7	S_8	S_9	S_{10}	···	S_{14}
H_1	0	0	0	0	0	0	0	0	0	0.230	···	0
H_2	0	0	0.203	0	0	0	0	0.013	0	0	···	0
H_3	0	0	0	0	0	0	0	0.001	0	0	···	0
H_4	0	0	0	0	0	0	0	0	0	0	···	0.151
H_5	0	0	0	0	0.212	0.014	0	0.003	0	0	···	0
H_6	0.176	0	0.003	0	0	0	0	0.045	0	0	···	0
H_7	0	0.002	0	0	0	0	0	0.009	0.219	0	···	0
H_8	0	0	0	0	0	0	0	0	0	0.002	···	0
H_9	0	0.204	0	0	0	0.001	0	0.028	0.002	0	···	0
H_{10}	0	0	0	0	0	0.001	0	0.006	0	0	···	0

Table 5. Semantic variable vectors of sentences using NMF

4.2 Extraction of Sentences

4.2.1 Document Summarization Using NMF

Lee et al. [22] proposed a novel method to select sentences based on NMF and defined (GRS) for the j^{th} sentence as follows:

$$\text{GRS} = \sum_{i=1}^{k}(H[i,j] \times \text{weight}(H[i,*])) \tag{5}$$

$$\text{weight}(H[i,*]) = \frac{\sum_{q-1}^{n} H[i,q]}{\sum_{p-1}^{r}\sum_{q-1}^{n} H[p,q]} \tag{6}$$

where the weight $(H[i,*])$ is the relative relevance among those i^{th} semantic features. It is clear that the generic relevance could reflect the major topics of sentences using representations of their semantic features.

Table 6 shows the sentence extraction process from the semantic variable vector $H[i,*]$ for sentences obtained by NMF in Table 4. By using Equation (5), sentence S_6 in Table 1 is extracted, which corresponds to the highest GRS (0.194).

Sentence	S_1	S_2	S_3	S_4	S_5	S_6	S_7	S_8	S_9	S_{10}	···	S_{14}
GRS	0.025	0.053	0.071	0.008	0.129	0.194	0.075	0.039	0.052	0.168	...	0.011

Table 6. Sentence extraction processing using GRS

5 EXPERIMENT AND EVALUATION

In this study, to reduce the size of a matrix A and to speed up computation of the NMF method this evaluation has no stemming process and no filtering of stop words, and the generic impressive expressions will be proposed.

5.1 Experiment and Evaluation

Table 7 shows impression dictionaries which are automatically constructed by using six impressive word dictionaries *P1*[4], *P2*[5], *P3*[6], *P4*[7], *P5*[8], *P6*[9] and one negative impressive word dictionary *N1*[10]. After constructing this dictionary, two Ph. D. students have checked and appended more significant impressive expressions.

Impression Dictionaries Word	Frequency	Impression	Original Word		
new(P2:new)(P5:new)(P6:new)	2 536	ps	newness	newness	new(P2)(P5)(P6)
time(P1:timely)(P5:timely)	1 337	pw	timely(P1)(P5)	timely(P1)(P5)	timed
percent	1 215	*es	percent (es)		
report(P4:reported)	1 202	pw	reporter	reporting	report's
like(P5:like)(P5:liking)(P6:liked)	1 155	ps	liked	likeness	like
offici(P6:officious)	982	*pw	officially	officially	official (pw)
because	977	*cw	because (cw)		
problem	476	*ns	problem (ns)	problems	
forc(P6:forced)(P6:forceful)	383	nw	forced(P6)(N1)	force	forceful
medic(P6:medicated)	322	*pnw	medication	medical (pnw)	medically
district	318	*nw	districts	district's	district (nw)
period	135	*ew	Periods(ew)	Per-iodizing	periodic

Table 7. Construction of impressive expression dictionary

In column 1 of Table 7, new (P2:new)(P5:new)(P6:new) means that the impressive expression "new" is matching with the impressive dictionaries, P2, P5 and P6, column 2 is representing the frequencies of expressions form all text corpus, and *pw means this impressive word picked up from the impression dictionary that has been built manually; for example, the word "official" appended to the exited dictionary. Table 8 shows the total number of impressive expressions automatically and manually.

Automatically strong and weak impressive expressions	Number	Manually strong and weak impressive expressions	Number
ps	352	*ps	645
pw	217	*pw	137
ns	171	*ns	262
nw	79	*nw	52
pnw	5	*pnw	388
es	22	*es	19
ew	16	*ew	18
cw	9	*cw	4
cs	11	*cs	2

Table 8. Total numbers of impressive expressions

[4] `http://www.the-benefits-of-positive-thinking.com/list-of-positive-words.html`

[5] `http://www.winspiration.co.uk/positive.htm`

[6] `http://www.creativeaffirmations.com/positive-words.html`

[7] `http://blog.emurse.com/2007/02/08/complete-list-of-english-power-words/`

[8] `http://www.mindmapinspiration.co.uk`

[9] `http://www.creativeaffirmations.com/positive-words.html`

[10] `http://eqi.org/fw_neg.htm`

Column 2 of Table 8 represents the number of the strong and weak impression expressions automatically decided while column 4 represents strong and weak impression expressions appended to the original dictionaries manually. In this study, the numbers of all impressive expressions are 5 722.

5.2 Data Collection and Evaluation System

- Test data
  Effective test data is using documents from DUC2006 data set [23] that can compare manual summaries by experts with the automatic summaries of the proposed approach. DUC2006 data set includes 50 documents as test data selected randomly. This test data includes 50 topics with 25 documents related to each topic [24]. Each document consists of segmented sentences and it has manual summaries up to 250 words. For DUC2006 data set, ROUGE evaluation systems [34] can compare generated summaries by the proposed method with manual summaries.

5.3 Size of Matrix

- Total number of words in matrix A is described as follows:
  - ANW: average number of words for each document
  - MXNW: maximum number of words for each document
  - MNNW: minimum number of words for each document.

Table 9 shows details of total number of words with frequency in matrix A.

	The proposed method			*The original methods (NMF)*		
	Keyword	Key Freq	Matrix Size	Keyword	Key Freq	Matrix Size
ANW	54.16	79.76	813.74	175.98	329.94	2 741.64
MXNW	121	193	2 783	407	789	8 954
MNNW	10	10	50	72	95	360

Table 9. Number of keywords in a matrix A for the original and the proposed methods

From Table 9, it clear that ANW of keywords for each documents by the proposed methods is 54.16 while it is 175.98 for the original which means that around 80 % of the keywords are filtering. At the same time, the matrix size could reduce by the same average.

Figure 2 shows the comparison of the matrix size of the proposed and original methods.

From results of Figure 2, it turns out that the proposed method can achieve about 80 % reduction of matrix A for the 100 KB text data. This indicates that, by using impressive expressions, the proposed method provides better performance in identifying summarization documents.

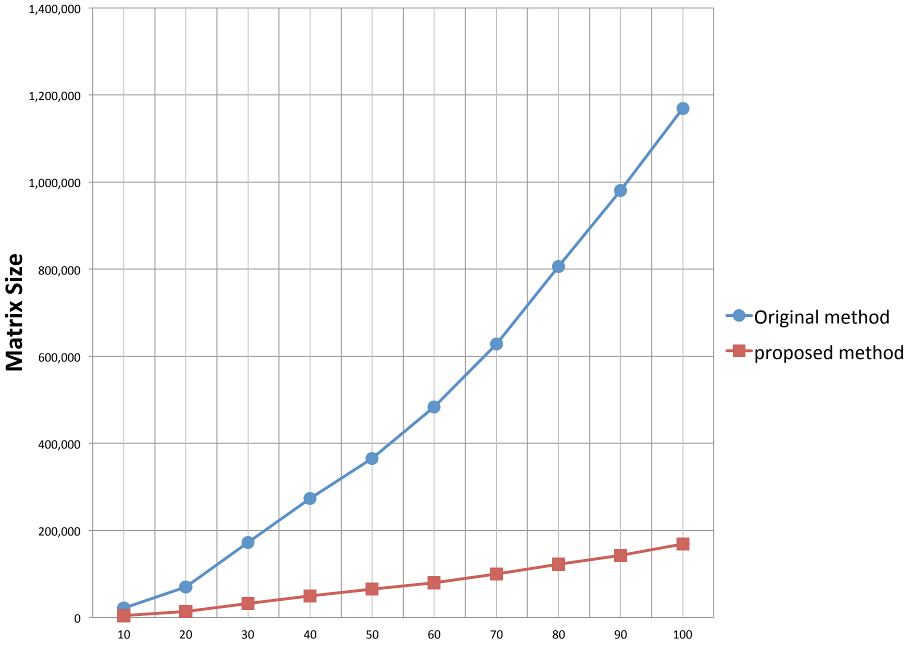


Fig. 2. Matrix size comparison for the proposed and original methods

5.4 Speed Experiments

The proposed system has been developed under Linux Ubuntu 10.04.2 LTS 64bit and 8 CPU of Intel Xeon W3520 (2.67 GHz) with 6 GB main memory. Figure 3 shows the speeding up of the computation of the NMF method for both original and proposed methods.

From the 100 KB text data of simulation results of Figure 3, it turns out that the time of the proposed method using impressive expressions is about 7 times faster than the original one.

5.5 Evaluation System Accuracy

This paper focuses on computing precision only, because we could not adjust the size of the whole documents and the work based on relevancy and speeding up.

5.5.1 Evaluation System

ROUGE scores were computed by running ROUGE-1.5.5 [34] with no stemming and no removal of stop words. The input file implemented so that scores of systems and humans could be compared.

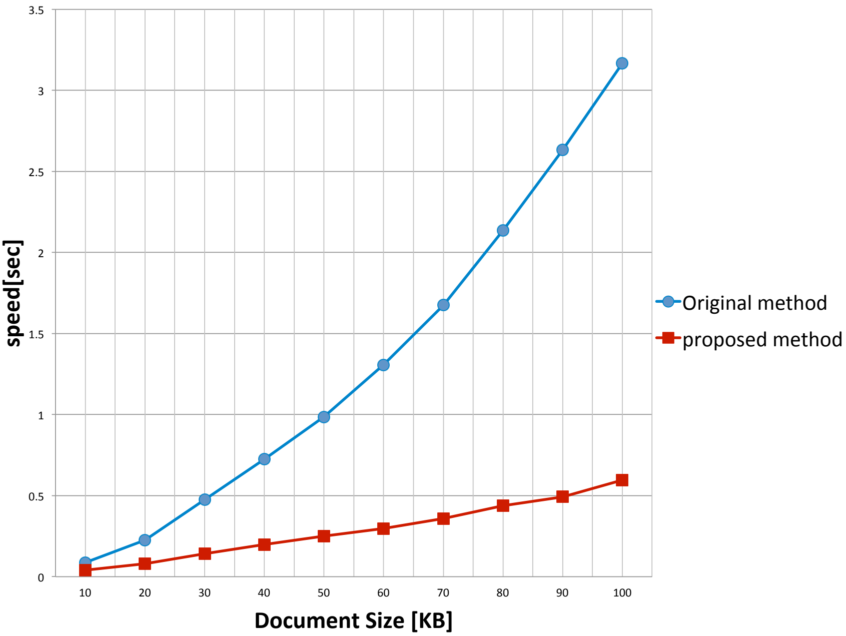


Fig. 3. Speeding up of the computation of the NMF method for original and proposed methods

ROUGE evaluation systems are used to compute precision by using ROUGE_N which represents precision between generated summary of the proposed system and manual summary. Let n be the length of the n-*gram*, $gram_n$ is the maximum number of n-*gram* in the generated summary and $Count_n(gram_n)$ is a set of manual summary.

$$\text{ROUGE-N} = \frac{\sum_{S\in\{\text{manualsummary}\}}\sum_{\text{gram}_n\in S} Count_n(\text{gram}_n)}{\sum_{S\in\{\text{manualsummary}\}}\sum_{\text{gram}_n\in S} Count(\text{gram}_n)} \tag{7}$$

In the system, five automatic evaluation methods are prepared in the ROUGE evaluation system ROUGE-N, ROUGE-L, ROUGE-W, ROUGE-S, and ROUGE-SU [34] as follows:

1) **ROUGE-N:** *N-gram co-occurrence statistics* which is a precision between a generated summary of the proposed system and manual summary.

2) **ROUGE-L:** *Longest Common Subsequence (LCS)* which compares similarity between two documents in automatic summarization evaluation.

3) **ROUGE-W:** *Weighted Longest Common Subsequence (WLCS)* which is called weight algorithm to assign different credit to consecutive in sequence matches.

4) **ROUGE-S:** *Skip-Bigram Co-Occurrence Statistics* which measures the overlap of skip-bigrams between a generation translation and manual translation.

5) **ROUGE-SU:** *Extension of ROUGE-S* which is obtained from ROUGE-S by adding a begin-of-sentence marker at the beginning of generated and manual sentences.

Table 10 shows comparison between precision of the original and the proposed NMF method using ROUGE evaluation; (N) means using frequency of occurring term in sentence only (No weight) while (B) means using binary weight i.e. the weight of term equal 1 if it is appears at least once in the sentence; otherwise the weight will equal 0, where 30 % means that the original method is using same dictionary of the proposed methods and proposed method with only *ps*, *ns* means that the proposed method is using only strong positive and negative impressive expressions, not all impressive expressions.

Precision evaluation for the original and the proposed (NMF) methods				
ROUGE evaluation	ROUGE-1	ROUGE-L	ROUGE-W	ROUGE-SU
Proposed method (N)	0.4180	0.37009	0.23015	0.18341
Proposed method only ps and ns(N)	0.39878	0.35775	0.2285	0.17137
Original method (N)	0.39625	0.35112	0.21682	0.16956
Original method no impressive expressions (N)	0.39068	0.34961	0.21603	0.1687
Original method (30 %) (N)	0.39368	0.35231	0.22016	0.17082

Table 10. Precision for the original and the proposed NMF method using frequency (N)

From evaluation results of Table 10, it is clear that the precision of the proposed method using (B) is somewhat higher than the original method among all ROUGE measures. Table 11 shows improvement in the precision of the original NMF method after using frequency.

Precision evaluation for the original and the proposed (NMF) methods				
ROUGE evaluation	ROUGE-1	ROUGE-L	ROUGE-W	ROUGE-SU
Proposed method (B)	0.41173	0.35922	0.2267	0.18143
Proposed method with only ps and ns (B)	0.41128	0.36545	0.22938	0.1784
Original method (B)	0.40892	0.36323	0.22554	0.17799
Original method not impressive (B)	0.39349	0.34799	0.22005	0.16856
Original method (30 %) (B)	0.40905	0.36398	0.22917	0.18046

Table 11. Precision for the original and the proposed NMF method using binary weight (B)

From Table 11, in the precision evaluation results, the weight (N) of the proposed method showed a better performance than the original method among all ROUGE measures. This means that using impressive word could affect the improvement of the original method precision.

5.5.2 Accuracy Evaluation

Table 12 shows the accuracy of sentence ranking using impressive expressions for the original and the proposed methods using 50 documents from DUC corpus.

From results of Table 12, it is clear that the accuracy of sentence ranking of the proposed method is higher than that of original methods by 8 %, 18 %, 46 %, 8 % for

Sentence Ranking Evaluation for the proposed and the original method				
ROUGE evaluation	ROUGE-1	ROUGE-L	ROUGE-W	ROUGE-SU
Accuracy Sentence Ranking Over (worst-case)	30 (25)	29 (25)	34 (19)	29 (24)
Accuracy Sentence Ranking Down (worst-case)	19 (8)	19 (8)	16 (10)	21 (9)
Accuracy Sentence Ranking Equal (worst-case)	1 (17)	2 (17)	0 (21)	0 (17)

Table 12. Accuracy of sentence ranking using impressive expressions

Sentence Ranking Evaluation for the proposed and the original method				
ROUGE evaluation	ROUGE-1	ROUGE-L	ROUGE-W	ROUGE-SU
Accuracy Sentence Ranking Over (worst-case)	26 (19)	29 (17)	36 (12)	26 (18)
Accuracy Sentence Ranking Down (worst-case)	22(7)	20 (6)	13 (7)	22 (9)
Accuracy Sentence Ranking Equal (worst-case)	2 (24)	1 (27)	1 (31)	2 (23)

Table 13. Accuracy of sentence ranking without impressive expressions

all ROUGE measurements, respectively. The proposed method indicates the best performance of the accuracy with ROUGE-W.

Table 13 shows the accuracy of sentence ranking without impressive expressions for original and proposed methods using the same documents.

From results of Table 13, it is clear that the accuracy of sentence ranking of the proposed method is higher than that of original methods by 22 %, 21 %, 36 %, 16 % for all ROUGE measurements, respectively. The proposed method indicates the best performance of the accuracy with ROUGE-W.

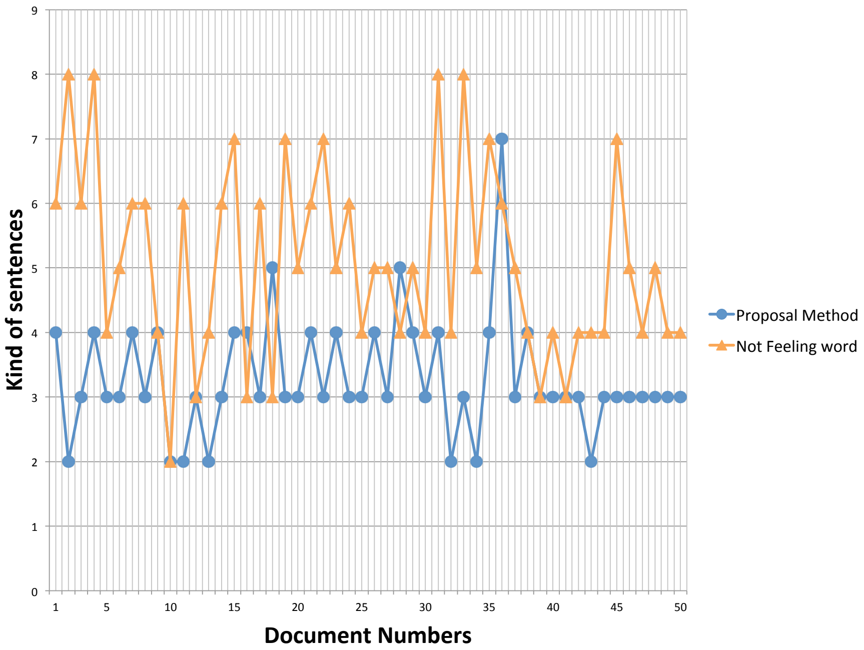


Fig. 4. Kind of extracted sentences with and without using impressive expressions

From the results of Figure 4, it turns out that our proposed method is stable for extracting sentences for each document summary by using impressive expressions, while there is a big scattering of extracted sentences for each document summary without using impressive expressions. This means that the proposed method could identify the kind of extracting sentences more successfully than the original one.

In conclusion, from the 100 KB text data of previous results, it turned out that the matrix size could be reduced by about 80 % and the speeding up of the computation of the NMF method becomes 7 times faster than with the original method without degrading relevancy of extracted sentences.

6 CONCLUSION

Although techniques using latent semantic analysis (LSA) and non-negative matrix factorization (NMF) have been applied to determine topics of documents, there are no researches on the reduction of the matrix and the speeding up the computation of the NMF method. In order to achieve this scheme, this paper has utilized the generic impressive expressions of newspaper to extract important sentences as the summary. Therefore, it has no stemming processes and no filtering of stop words. This paper has proposed an impression-based summarization scheme of newspapers because they have widespread in the world. The proposed method has introduced impression expressions for newspapers and their measurements are applied to the NMF. From 100 KB text data of experimental results, it turns out that the matrix size is smaller by 80 % and the speeding up of the computation of the NMF method becomes 7 times faster than in the original method, without degrading the relevancy of extracted sentences.

Future works could extend this work using large corpus to get more speeding up of the computation and to increase the relevancy of extracted sentences.

REFERENCES

[1] Erkan, G.—Radev, D. R.: : Lexrank: Graph-Based Centrality as Salience in Text Summarization. J. Artif. Intell, 2004, Res. 22, pp. 457–479.

[2] Ding, C.—Li, P. T.—Park, H.: Orthogonal Nonnegative Matrix Trifactorizations for Clustering. In Proceedings of SIGKDD 2006.

[3] Reeve, L. H.—Han, H.: The Use of Domain-Specific Concepts in Biomedical Text Summarization. Information Processes and Management, Vol. 43, 2007, pp. 1765–1776.

[4] Wang, D.—Zhu, S.—Li, T.—Chi, Y.—Gong, Y.: Integrating Clustering and Multi-Document Summarization to Improve Document Understanding. In Proceedings of CIKM 2008.

[5] Chang, T. M.—Hsiao, W. F.: A Hybrid Approach to Automatic Text Summarization. IEEE International Conference 2008, pp. 65–70.

[6] Chongsuntornsri, A.—Sornil, O.: An Automatic Thai Text Summarization Using Topic Sensitive Page Rank. International Symposium on Communications and Information Technologies (ISCIT '06) 2006, pp. 547–552.

[7] Hennig, L.—Umbrath, W.—Wetzker, R.: An Ontology-Based Approach to Text Summarization. Proc. of EEE/WIC/ACM International Conference onWeb Intelligence and Intelligent Agent Technology 2008, pp. 291–294.

[8] Jing, H.: Sentence Reduction for Automatic Text Summarization. Applied Natural Language Conferences, Proceedings of the Sixth Conference on Applied Natural Language Processing, Seattle, Washington, USA 2000, pp. 310–315.

[9] Kruengkrai, C.—Jaruskulchai, C.: Generic Text Summarization Using Local and Global Properties of Sentences. In Proceedings of the IEEE/WIC international Conference on Web Intelligence (IEEE/WIC '03) 2003.

[10] Uzeda, V. R.—Pardo, T.—Nunes, M.: Evaluation of Automatic Text Summarization Methods Based on Rhetorical Structure Theory. 8th International Conference on Intelligent Systems Design and Applications (ISDA '08) 2008, pp. 389–394.

[11] Atlam, El S.—Fuketa, M.—Morita, K.—Aoe, J.: Similarity Measurement Using Term Negative Weight and its Application to Word Similarity. Information Processes and Management, Vol. 36, No. 6, pp. 717–736.

[12] Alani, H.: Automatic Extraction of Knowledge from Web Documents. In Workshop of Human Language Technology for the Semantic Web and Web Services, 2nd International Semantic Web Conference, Sanibel Island, Florida, USA 2003.

[13] Bennett, K. P.Campbell, C.: Support Vector Machines: Hype or Hallelujah. SIGKDD Explorations, Vol. 2, 2000, No. 2, pp. 1–13.

[14] Jing, H.—McKeown, K. R.: Cut and Paste Based Text Summarization. ACM International Conference Proceedings, Series 4, 2000, pp. 178–185.

[15] Liang, S. F.—Devlin, S.—Tait, J.: Investigating Sentence Weighting Components for Automatic Summarization. Information Processes and Management, Vol. 43, 2007, No. 1, pp. 146–153.

[16] Mani, I.: Automatic Summarization. John Benjamins Publishing Company 1999.

[17] Catanzaro, B.—Sundaram, B.—Keutzer, K.: Fast Support Vector Machine Training and Classification on Graphics Processors. In International Conference on Machine Learning 2008.

[18] Yoshinari, T.—Atlam, El S.—Morita, K.—Kiyoi, K.—Aoe, J.: Automatic Acquisition for Sensibility Knowledge Using Co-Occurrence Relation. IJCAT, Vol. 33, 2003, No. 2/3, pp. 218–225.

[19] Zha, H.: Generic Summarization and key Phrase Extraction Using Mutual Reinforcement Principle and Sentence Clustering. In Proceedings of the 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '02), Tampere, Finland 2002, pp. 113–120.

[20] Yeh, J.-Y.—Ke, H.-R.: Text Summarization Using a Trainable Summarizer and Latent Semantic Analysis. Information Processes and Management, Vol. 41, 2005, pp. 75–95.

[21] Li, W.—Li, B.—Wu, M.: Query Focus Guided Sentence Selection Strategy for DUC 2006. In Proceedings of Document Understanding Conference (DUC '06) 2006.

[22] Lee, J.-H.—Park, S.—Ahn, C.-M.—Kim, D.: Automatic Generic Document Summarization Based on Non-Negative Matrix Factorization. Information Processing and Management 45, 2009, pp. 20–34.

[23] Duc' 06, `http://duc.nist.gov/`.

[24] Hao, T. D.: Overview of DUC 2005. In Poceedings of the Document Understanding Conference (DUC '05) 2005.

[25] Willian, B. F.—Ricardo, B.-Y.: Information Retrieval: Data Structure & Algorithms. Prentice Hall 1999.

[26] Tiedens, L. Z.: Feeling low and Feeling High: Associations Between Social Status and Emotions. Unpublished doctoral dissertation, University of Michigan, Ann Arbor 1998.

[27] Tiedens, L. Z.: Anger and Advancement Versus Sadness and Subjugation: The Effect of Negative Emotion Expressions on Social Status Conferral. Journal of Personality and Social Psychology, Vol. 80, 2001, No. 1, pp. 86–94.

[28] Adler, R. S.—Rosen, B.—Silverstein, E. M.: Emotions in Negotiation: How to Manage Fear and Anger. Negotiation Journal, 1998, pp. 161–177.

[29] Kadoya, Y.—Morita, K.—Fuketa, M.—Oono, M.—Atlam, El S.—Sumitomo, T.—Aoe, J.: A Sentence Classification Technique by Using Intention Association Expressions. International Journal of Computer Mathematics, Vol. 82, 2005, No. 7, pp. 777–792.

[30] Lee, D.—Seung, H.: Learning the Parts of Objects by non-Negative Matrix Factorization. Nature 401, 1999, pp. 788–791.

[31] Liu, S.: Enhancing E-Business-Intelligence-Service: A Topic-Guided Text Summarization Framework. Seventh IEEE International Conference on E-Commerce Technology (CEC) 2005, pp. 493–496.

[32] Lee, D.—Seung, H.: Algorithms for Non-Negative Matrix Factorization. Advances in Natural Information Processing Systems, 13, 2001, pp. 556–562.

[33] Wild, S.—Curry, J.—Dougherty, A.: Motivating Non-Negative Matrix Factorization. In Proceedings of the 26th Annual International ACM SIGR Conference on Research and Development in Information Retrieval, Toronto, Canada 2003, pp. 267–273.

[34] Lin, C. Y.: ROUGE: A Package for Automatic Evaluation of Summaries. In Proceedings of Workshop on Text Summarization Branches Out, post-conference workshop of ACL 2004.

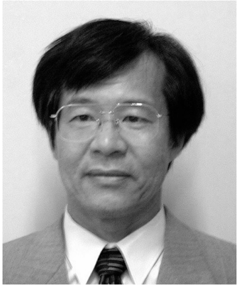
Abdunabi Ubul received his B. Sc. degree in economics and management information from Xinjiang University, China in 2004. He has received his M. Sc. degree from Department of Economics, Faculty of Integrated Arts and Sciences, University of Tokushima, Japan in 2008. He is now a Ph. D. student at the Department of Information Science and Intelligent Systems. His research interests include information retrieval, natural language processing and document processing.

Hiroya Kitagawa received his B. Sc. and M. Sc. degrees in information science from University of Tokushima, Japan in 2008 and 2010, respectively. He is now a Ph. D. student at Department of Information Science, University of Tokushima, Japan. His research interests include information retrieval, natural language processing and document processing.

El-Sayed Atlam received his B. Sc. and M. Sc. degrees in mathematics from Faculty of Science, Tanta University, Egypt in 1990 and 1994, respectively, and his Ph. D. degree in information science and intelligent systems from University of Tokushima, Japan, in 2002. He has been awarded Japan Society of the Promotion of Science (JSPS) Postdoctoral Fellow from 2003 to 2005 in Department of Information Science & Intelligent Systems, Tokushima University. He is currently Associate Professor at the Department of information Science and Intelligent Systems at University of Tokushima, Japan. 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 of the Egyptian Mathematical Association (EMA). His research interests include information retrieval, natural language processing and document processing.

Masao Fuketa received his B. Sc., M. Sc. and Ph. D. degrees in information science and intelligent systems from University of Tokushima, Japan in 1993, 1995 and 1998, respectively. Between 1998 and 2000 he was a research assistant in information science and intelligent systems at University of Tokushima, Japan. He is currently a researcher in the Department of Information Science & Intelligent Systems, Tokushima University, Japan. He is a member in the Information Processing Society in Japan and The Association for Natural Language Processing of Japan. His research interests include sentence retrieval from huge text databases and morphological analysis.

Kazuhiro Morita received his B. Sc., M. Sc. and Ph. D. degrees in information science and intelligent systems from University of Tokushima, Japan, in 1995, 1997 and 2000, respectively. Since 2000, he was research assistant at the Department of Information Science & Intelligent Systems, Tokushima University, Japan. His research interests include sentence retrieval from huge text databases, double-array structures and binary search trees.



Jun-ichi Aoe received his B. Sc. and M. Sc. degrees in electronic engineering from the University of Tokushima, Japan, in 1974 and 1976, respectively, and his Ph. D. degree in communication engineering from the University of Osaka, Japan in 1980. Since 1976 he has been with the University of Tokushima. He is currently a Professor in the Department of Information Science & Intelligent Systems, Tokushima University, Japan. 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 trie compaction algorithms for large key sets. He is the 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 Natural Language Processing of Japan and the IEEE Computer Society