Res. Rep. Fac. Eng. Mie Univ., Vol. 19, pp. 191-198(1994)

Review

Information Extraction from Electronic Mail

Atsuo KAWAI, Takeyuki TSUKAMOTO, Tsutomu SHIINO
(Department of Information Engineering)

(Received September 16.1994)

Abstract

The former researches on information extraction are only from documents with only sentences. But many documents include non-sentence areas, i.e., tabular areas such as itemizations, tabular forms and marsharing of nouns. This paper describes an information extraction method from tabular areas. The new algorithm consists of three steps. First, tabular areas are recognized by using a 2-dimensional arrangement of letters, blank spaces, parts of speech, and semantic features of nouns. Second, the tabular areas are subdivided into blocks. And last, information extraction is done by matching the block against the frame. Words themselves and semantic feature of words are the clues clues to fill in the slots in the frame. Desk simulation for electronic mail of computer sale information shows a precision rate of 83%, 92% and 78% for each step.

Keywords: information extraction, tabular forms, itemizations, document structure, 2-dimensional arrangement

1.Introduction

The technique to draw out only the specified information, i.e., the technique of contents extraction, is an indispensable technique for adjustment of document information and automatic construction of a data base from document data of large quantity. Key word extraction from a text is the first step of contents extraction from document data. With regard to key word extraction, for example, there are various researches from the view points of how to decide an important key word. But, relation between key words is unknown, and the informaion which was gained is not changed structure in only mere key word extraction. In that place, it changes structure it is thought as following step of contents extraction to take out extraction of a key word extraction of a key word in other words related information between key words. Relation between a key word can be expressed by a grammatical function such as a Japanese particle,

a verb, ... (example : prices of main part and floppy disk equipment is 15 ten-thousand, 3.5 ten-thousand, respectively). Therefore, affecting and receiving it of a sentence, they perform structure and a grammatical analysis between sentences, so that can get relation between key words. Research of contents extraction from the ordinary sentence which provided a particle, the verb and so on which did it in this manner has been performed intending for learned paper [1], patent [2], introduction article [3][4] of a product, a document of [5] and so on newspaper.

But it is as scarce a number in a novel as in the document that we turn into an everyday eye to be composed in only the sentence that all a document provided a particle, a verb and so on. On the contrary, the itemized statement table, next composition element besides a sentence that called an enumeration of a word is implied in a document in lots of documents besides it. Even though this is ignored, it possible in the document that table, the itemization which is a composition element besides a sentence are used in support to perform contents extraction. But, when the whole is formed of itemization and only table, it is numerous in a document in an electronic mail and a businesslike document. To such document, the former technique that relation between key words is extracted based on grammatical relation within a sentence between sentences can not be applied.

Special knowledge of a descriptive object field is strengthened, so that moreover lots of systems besides grammatical information have gone amount accuracy of management change. But, there is also the indication it requires for the massive amount of work severely in comparison with a case of a data base, and an enough knowledge base is made every [6], special field, and it is difficult as for expansion of knowledge to deal with a text base it is adjusted to movement of the world, and to maintain insufficient knowledge base is used, it changes structure the technique that extraction of a key word is possible needs to be introduced to a document including an element besides a sentence.

Structural recognition of a document in other words table and scope of itemization are understood, and by it tells and is settled about an individual item, and this scope is diviced into (1 - of number go become), so that, the way relation between key words is understood is proposed in this paper in that place. Positioning of this research was mainly expressed from natural language management and a point of view of contents extraction in the above. The itemized statement next management process that document structure of table is understood is implied in the formula that they suggest in this paper next. About this point, positioning of research is expressed.

About the itemized statement next research which understands document structure of table in respect of 2 pieces.

One is researches [7] from the viewpoint of logical structural description of a document of ODA and SGML. About these research, recognition of reference structure to table is going automatically from recognition of itemization and description in paragraph. But, table is grasped as a blackbox, and has not turned into the research which understands inside structure of table.

Another is portrait recognition and a field on an extension line of character recognition by research field [8][9][10] which is soon called so called document portrait understanding. In a field of document portrait understanding, they are going at recognition of inside structure of table. But, only specific table structure which was fixed on a document manages these most research only. Moreover, though there is not been also the research [8] which understands diverse table structure, it is not corresponding to recognition of table of a veil vote, itemizati and diverse form structure of mixture of table in part. Moreover, since input is portrait data, a

horizontal - perpendicular frame ruled line is also input as information, and it has resulted in indispensable information in recognition of table structure either in a field of document portrait understanding. To this, since input is a document of character data (character code), though a horizontal - perpendicular frame ruled line exists on logic, this research is not printed on the document which is output. For this reason, there is the difference that they must guess based on the meaning that 2 dimensions stationing of a character string and a word have the frame ruled line which exists on a fiction.

When a human being understands document structure, moreover research in respect of 2 pieces which was mentioned in a top is not using the information which is meaning of a word and the natural language management of grammar which is used either. Moreover, it researches and is not accompanied from the viewpoint of contents extraction.

2.Contents extraction method

2.1 Document class treated in this paper

The document that deal is turned into the document which is formed of the itemized statement a sentence, next table, which is expressed by an enumeration of a word, a character string in this paper.

Following the way to mention it by 1., former contents extraction has been performed based on an analytic result of a sentence mainly. To this they analyze document structure based on 2 dimensions (physical) stationing character and so on, and is proposing the technique that contents extraction is performed based on this research. In that place, 2 dimensions stationing character and so on in a document has meaning, an every part of a document for is classified on 3 pieces not to hold in this paper.

(1)paragraph : 2 dimensions stationing of character and so on does not have meaning. It is composed from only a sentence. A research object of former contents extraction.

(2)tabular form: 2 dimensions stationing of character and so on in other words layout information has meaning. It consists of a sentence and an enumeration of a word. So called itemization is also included in this inside.

(3)others: (1), (2) besides part, for example, noun of enumeration by exist, the 2 dimensions stationing in terms of meaning at have).

It is especially mentioned about contents extraction techniques from a table part in this inside, this paper. In addition "paragraph", a term of "tabular form" are used in the meaning which was defined by the above (1), (2) by the following description.

2.2 Overall flow chart

Following the way to mention it by 1., contents extraction from a table part and a formula of contents extraction from paragraph differ. For this reason, all a document is divided into paragraph and a table part as shown in Fig. 1 first of all. Moreover a stream of whole management is shown in Fig. 2. In other words, it is managed in the order of the following.

(1)A morphemic analysis is performed.

(2)table all part is understood in the unit of a line.

(3)table part is divided every line with concluding. As a result of being divided, it goes and become the number of 1-. In this paper, they call this a block.

(4)By applying it to a frame, they perform related recognition between the words which was described within block.

It is mentioned minutely about the rule that it is used by (2) - (4) less than.

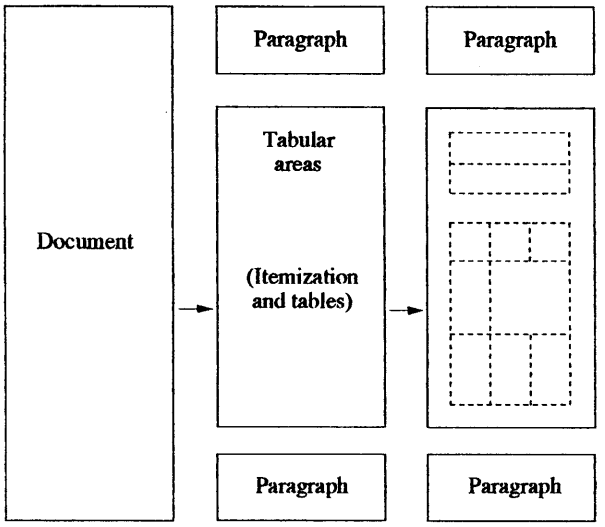

Fig. 1 Partition of a document

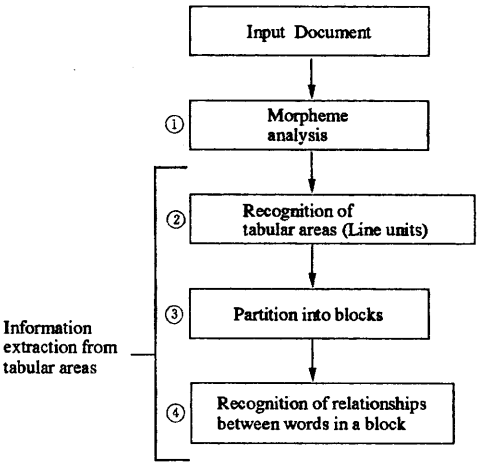

Fig. 2 Overall flow chart

2.3 Recognition of tabular form areas

Recognition of scope of a table part is going using a character, a part of speech and a characteristic of 2 dimensions stationing of a meaning attribute. A table part is divided to classify it roughly into itemization and table. Therefore, the recognition rule which respectively corresponded needs and become. The document example that each rule are applied is shown below.

2.3.1 Recognition rules of itemization

i)Numeral and so on of identity exist similar row (line of head on limit).

The following product is sold.
1.PC-9801 DA2 NEC 80,000 yen.
2.Monitorial (XC-1498C2) Mitsubishi 25,000 yen.
3.Memory (PIO-DA134) IO data 10,000 yen.

2.3.2 Recognition rules of table

i)Special sign (: ;) of identity or blank parts exists on a same row.

EPSON : PC-386NARX.
Logitec : 80MBHDD.
AIWA : PV-A24V5.
Besides there is necessary attachment goods everything.

ii)Table part is enclosed with the line which is formed of only special signs.

The following product is sold.

****

Main part.	NEC.	PC9801FA2.
Monitor.	Mitsubishi.	XC-1498C2.
Hard disk.	ICM.	HC-100ES.
Memory.	Melco.	EMJ-4000L.

****

We concede at hundred thousand yen , a full set.

iii)Indentation (head row of plural lines exists from a head line of a phrasal section in the right side).

iv)Part of speech composition of a table part.

(1)It is composed from noun (a part of inflection)

(2)auxiliary verb, a particle, a business remark, included.

v)There are the words which has a similar meaning attribute within identical table.

2.4 Block division

2.4.1 Block division of itemization.

i)A numeral of a head row and special sign. Example document is shown in 2.3.1

2.4.2 Block division of table.

i)Special signs. Example document is shown in 2.3.2

ii)Blank parts (white spaces) Example document is shown in Fig.4

iii)Default processing

A table part by 2.3, a special sign and the blank part and so on that it becomes a trace of the block division which was shown in a top do not sometimes exist. Then it is managed as 1 line 1 block.

2.5 Recognition of relation between words in the block.

It corresponds 1 piece of block(jded by 2.4. and, they decide for a letter surface of each noun which exists in a block and a meaning attribute, which slot in a frame to compensate.

For example 1 block is corresponding to a frame of description of 1 table of articles of main part /accessories (name, type turn, maker name, price, specification, class) in a document of secondhand goods transactional information of the personal computer which is mentioned by 3.This is shown in Fig.3

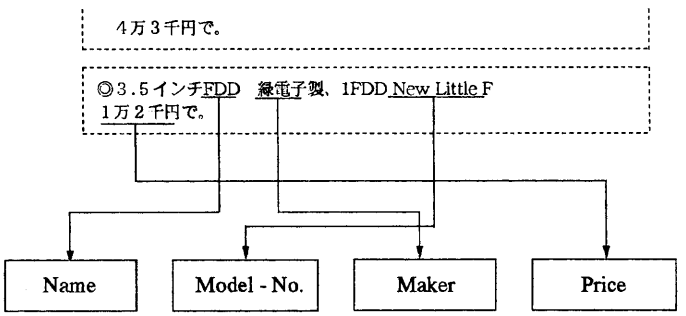

Fig. 3 Recognition of relationships between words in a block
(Dotted line represents a block)

It differs from the case that the item that it should be written beforehand is appointed on precision in a frame ruled line like a format of a veil vote in time of table and the itemization which exists in a document. Therefore, when than (slot) item words which should be

extractedare included, an entry eye sometimes is insufficient. When being insufficient, the necessity that the slot which is not buried in a frame is reasoned is caused. For example, it is the case that name (personal computer main part) and maker name (NEC) are reasoned from description of only type turn (PC9801RA). Knowledge about an object field needs and turns into reasoning. It has been realized from the high knowledge of mounting effect which performs complete reasoning.

3.Experiment and considerations

3.1 Experimental Documents

Following the way to mention it by 1., a document including a composition element besides the itemized statement table, next sentence of an enumeration of a word exists numerously in a document. For example, 70% of a page of association of the information processing academy which is thought a single kind of a businesslike document or more is occupied by the table, which is not enclosed with the itemized statement next frame ruled line. But, if they think by natural language and ratio g of comparison of a programming language, document structure of the publication that regulation about a form is done on the side of a writer corresponds to a programming language. In other words, analytic accuracy of system document structure decides whether a document structure analytic system reflects the specification which was done the regulation that a writer of 1 person (or 1 company) has how faithfully, and is hard to turn into an evaluation of a universal document structure analysis. Therefore, document structure of kinds is desirable to exist on spontaneous generation on an experimental object document.

In thinking from an aim of contents extraction itself, because the information which should be extracted must turn into a similar kind, moreover they need to intend for the document crowd that descriptive contents can limit to a certain measure of scope through an object all document.

Secondhand goods transactional information of computer network was noted from the point of view that the form of a document that the number the document that a descriptive object could limit to a certain measure of scope was opened to the public, and which was completed as an experimental object could receive was not controlled by the definite specification in that place. Concretely, a personal computer in notice board service of Nifty-Serve sold and bought it, it was noted.

27 FBH014433 8/03 34
Tsukamoto Takeyuki
PC9801RA21 main part.
110,000 yen.
120 mega HDD.
Midori-Denshi SCSI NOVA V120.
Like a new article after purchase 1 month.
Some applications are installed.
43,000 yen.
3.5 inch FDD Midori-Denshi New Little F.
12,000 yen.
Inner 4MRAM Melco EDA4000.
10,000 yen. A box and manual are equipped completely.
If you come home, I'll give you game software (slightly old). Else, it is sent by express delivery (dispatched free).
FBH01433 TSUKA.

Fig. 4 An example of electronic mail.

An electronic mail concerning secondhand goods transaction of a personal computer is shown in Fig. 4. A table part, paragraph become the parts that it is respectively shown in Fig. 4.

Moreover contents extraction results of Fig. 4 are shown in table 1. A goal of contents extraction was a main part of a personal computer and a name of an accessories, a model-No. a maker name, a price.

name	model - No.	maker	price
CPU	PC9801RA21	NEC	110,000 yen
Hard disk	NOVA V-120	Midori-denshi	43,000 yen
Floppy disk	New Little F	Midori-denshi	12,000 yen
Memory	EDA 4000	Melco	10,000 yen

Table 1 Information extraction from the document in Fig.4.

They manually removed it the document A which was described about the document which stuck and was written by the document which stuck and was written the document which overlapped and was contributed, only software, only a peripheral equipment, a machine type about 10 years front or more, and investigated 150 matters which was left out of the document which was recorded over several weeks, about 600 matters.

The document which consisted of this inside, only paragraph consisted of paragraph and a table part about 40% of the whole, the remaining 60% (88 matters), and turned this 88 matters an experimental object. As a whole tendency, when the number of the table of articles that we hope to sell is rare, it consists of only paragraph, and is also descriptive small in quantity. When the number of a table of articles is plentiful, main information summarizes and is carried on a table part frequently. The name which was turned into an extraction goal, a type turn, a maker name, information of a price of each table of articles are carried on a table part in the case of the most part in the document that paragraph and a tabular form mix together.

3.2 Evaluation.

The result of desk simulation of step 2.3, 2.4, 2.5 is

Table all partial recognition. 73/88 (83%).
Block division of a table part. 67/73 (92%).
Related recognition between words. 52/67 (78%).

The document example which failed in related recognition between words is shown below.

1.CPU NEC PC-9801 DA2.
(a box, accessories included).
2.Monitor XC-1498C2.
(a box included).
3.Memory IO DATA PIO-DA134(8M).
(main part on attach finished).
4.Others CYRIX CX486DLC33GP.
CYRIX. CX83D8733GP.
Printer ribbon (black, color).

It is not 1 table of articles, and is the reason why 3 lists of articles are described on the other blocks in this document.

4.Conclusion

They appeared frequently in the document which was described in Japanese, and suggested about the itemized statement the table that the technique that they dealt in a former natural language management technique did not exist, next technique which managed a document of an enumeration of a word in this paper. Management is promoted with a morphemic analysis, a sentence structure analysis, a meaning analysis, a contextual analysis, in the technical system which was established centering around former machine translation. Therefore, contents extraction from a class wrote it table, the article that a technique of a sentence structure analysis could not apply, and could not seek for a morphemic analytic level in other words key word extraction, relation between key words.

By nothing a character string in a document, 2 dimensions arrangement of a word and so on, they performed structure of a document change, and proposed the technique which fixed relation between key words in this paper. It can not say an analytic success rate of all management with high value sufficiently in 59% (88 matters in 52 matters), and can say a technique of contents extraction for a document including table to think not to exist an up to this time at all with the value that they can estimate.

There is the following point as a future problem. The extraction result which was got is looked up by concern of a user by the equal various forms only the product classification which is looked up from a product name is appointed, and that they look up from a price. They need to think about the data structure which can correspond to a reference demand of such kinds in other words housing form to a data base.

References

1)H.Inose,T.Saito,K.Hori:A Paper Abstract Understanding and Generation Assistance System by Means of Scenario,Trans.IPS Japan,Vol.24,No.1,pp.22-29(1983).
2)S.Takamatsu,K.Kusaka,H.Nishida:Automatic Extraction of Relational Information from Technical Abstracts,Trans.IPS Japan,Vol.25,No.2,pp.216-224(1984).
3)H.Matsuo:Information extraction method based on hierarchical matching of extraction patterns,IPSJ Technical Report,NL99,No.2,(1994).
4)E.Komatsu,Y.Kato,H.Yasuhara,T.Shiino:Summarizing Support System COGITO, IPSJ Technical Report,NL64,No.11,(1987).
5)P.Jacobs,L.Rau:SCISOR:Extracting information from on-line news,Comm.ACM, Vol.33,No.11,(1990).
6)K.Akiyama:The Issue of Intelligent Information Retrieval to Textbase,IPSJ Technical Report,DB64,No.3,(1988).
7)M.Doi,M.Hukui,K.Yamaguchi,Y.Takebayashi,I.Iwai:Development of Document Architecture Extraction,Trans.IEICE,Vol.J76-D-II,No.9,pp.2042-2052(1993).
8)Q.Luo,T.Watanabe,N.Sugie:Structure Recognition of Various Kinds of Table-Form Documents:Trans.IEICE,Vol.J76-D-II,No.10,pp.2165-2176(1993).
9)M.Yamada:Conversion Method from Document Image to Logically Structured Document Based on ODA,Trans.IEICE,Vol.J76-D-II,No.11,pp.2275-2284(1993).
10)A.Yamashita,T.Amano:A Model Based Layout Understanding Method for Document Images,Trans.IEICE,Vol.J76-D-II,No.10,pp.1673-1681(1992).