An Automatic Method of the Extraction of Important Words from Japanese Scientific Documents

Makoto NAGAO*, Mikio MIZUTANI* and Hiroyuki IKEDA*

Abstract

An automatic method is developed which extracts important words(typical words or possibly key words) from Japanese scientific documents.

The idea is to extract the words which appear frequently in some particular documents, but which appear seldom in other documents. To measure this particularity of word usage we utilized the idea of χ^2 test.

The experiment is done on a text-book of chemistry of the middle school, and "Current Bibliography on Science and Technology (Electrical Engineering)" edited by JICST. More than 120,000 Japanese words are handled and the results are fairly satisfactry.

1. Introduction

In the field of information retrieval, indexing to documents is an essential task to be done, and the indexing terms must be established. Usually the indexing terms are determined by human specialists in each specific field. We wanted to do this automatically by computer. We are concerned especially with documents in Japanese language, where the following special problems of Japanese language must be solved.

(1) Words are not separated by a space or something in a sentence. Segmentation of a sentential string into a sequence of words is one of the most difficult problems.

(2) Characters used are Katakana's, Hirakana's and Chinese Characters, the total of which amounts to a few thousand.

(3) Each Chinese Character expresses a certain meaning, and two or more of it are easily combined to form a new noun.

For the extraction of important words from documents, grammatical analysis of sentences is impossible, because it takes a long time and may be incomplete. We considered that important words in scientific literatures are usually nouns, and we attempted to

This paper first appeared in Japanese in Joho—Shori (Journal of the Information Processing Society of Japan), Vol. 17, No. 2 (1976), pp. 110~117.

* Department of Electrical Engineering, Kyoto University

extract nouns from sentential strings. Almost all the nouns of a specific scientific field are expressed either by Chinese characters or by Katakana characters. This helped us greatly for the extraction of nouns because we have only to extract character sequences of Chinese characters and Katakana's.

2. Algorithm to Extract Important Words

It is widely known that important words are those whose relative frequencies of appearance in documents are not small or large, but just middle. We modified this idea in the following way: (1) important words in a area will appear often in the specific area, but seldom in other areas, and (2) common words will appear equally in all the areas. To test this situation we employed χ^2(chi square) distribution.

We define some variables as follows. m is the number of different words. n is the number of all the different areas we are interested in. $x$ij is the frequency of word i in area j. $m$ij is the expected relative frequency of word i in area j. Then

$$m_{ij} = \frac{\sum_{j=1}^{n} x_{ij}}{\sum_{i=1}^{m} \sum_{j=1}^{n} x_{ij}} \times \sum_{i=1}^{m} x_{ij}$$

and

$$\chi_i^2 = \sum_{j=1}^{n} \frac{(x_{ij} - m_{ij})^2}{m_{ij}}$$

3. Extraction of ImportantWords in Chemistry Text-Book

We used a part of a middle school text-book of chemistry as an example. The chapters and sections are shown in Table 1 together with the area identifications which we set from the contents. In this case n=10. The number of nouns we treated by computer was 16363 words, in which different nouns were 1371. The word extraction results are shown in Tables 2 and 3. The results are fairly satisfactory. A problem is that the word "acid" was not picked up in area 4, because the concept of "acid" was used diversely as oxidation, oxide, acidity, acidify and so on, and the count of "acid" was not big. Another problem is, for example, that the concept of "the low of conservation of mass" is to be an important word, but that this concept is expressed by three words, each of which are not particularly important, and could not be picked up.

Table 1 Contents of chapters 8 and 9 of the textbook

Contents	area code	Contents	area code
Chapter 8 Material and Atoms		Chapter 9 Function of Electric Current	
Section 1 Chemical Compounds and Elements 1. Change of Materials 2. Resolution of Materials 3. Combination and Chemical Compounds	1	Section 1 Electric Current and Electric Voltage 1. Electric Current Which Flows in Small Lump 2. Function of Cell 3. Relation Between Electric Current and Voltage Which Flows in Metalic Wire	7
4. Change of Material and Molecule 5. Element and Atomic Symbol 6. Flame Reaction	2	4. Electric Registance of Metalic Wire 5. Electric Current Which Flows in Water Solvent 6. Property of Circuit 7. Direct Current and Alternating Current	8
Section 2 Chemical Change and Mass of Materials 1. Electrolysis of Water 2. Composition of Water 3. Change of Materials and Mass 4. Law of Conservation of Mass	3	Section 2 Electric Energy 1. Heat Generation by Electric Current 2. Electric Power	9
5. Oxidation of Metals 6. Reduction of Oxidized Metals and Law of Definite Proportions 7. Metals and Acidic Materials 8. Chemical Change and Energy	4	Section 3 Electron and Electric Current 1. Electric Current Which Flows in Gas 2. Cathode Rays and Electron 3. Diode and Electron 4. Electric Current and Electron Which Flows in Metals	10
Section 3 Chemical Change and Molecule, Atom 1. Smaller Particles than Molecule 2. Molecule and Atom of Chemical Compound 3. Size of Atom and Molecule	5		
4. Chemical Change and Combination of Atoms 5. Formula to Represent Materials 6. Formula to Represent Chemical Change	6		

Table 2 The 20 words selected according to χ^2-Value.

Rank by χ^2	Rank by Frequency	Word	Frequency	χ^2	Areas which show the peaks of χ^2
1	47	Atom	59	11.93	5,6
2	15	Electric Current	145	11.86	7,8,9,10
3	69	Element	34	11.85	2
4	25	Molecule	95	11.27	2,5,6
5	75	Sodium Chloride	31	10.30	1
6	98	Electron	23	10.14	10
7	112	Molecular Formula	18	10.09	6
8	89	Small Lump	26	9.75	7
9	84	Sodium Chlorate	28	9.69	1
10	82	Particle	29	9.12	1
11	149	Heater Wire	13	9.05	9
12	166	Electric Power	12	8.35	9
13	45	Express	61	8.31	6
14	108	Zinc	19	8.03	4
15	31	Voltage	81	7.99	7,8,9,10
16	19	Material	120	7.18	1,2,3,4,5,6,
17	78	Number	30	7.13	6
18	114	Dry Cell	17	7.11	7
19	124	Heat	16	6.80	9
20	162	Formula of Chemical Reaction	12	6.73	6

Table 3 The first 4 words selected from each area

area code	First Four Important Words
1	Sodium Chloride, Sodium Chlorate, Particle, Material
2	Element, Molecule, Material, Flame
3	Material, Hydrogen, Druggists' Scales, Chemical change
4	Zinc, Material, Hydrogen, Create
5	Atom, Molecule, Material, Combine
6	Atom, Molecule, Molecular Formula, Express
7	Current, Small Lump, Voltage, Dry Cell
8	Current, Voltage, Circuit, Metalic Wire
9	Current, Heater Wire, Electric Power, Voltage
10	Current, Electron, Voltage, Flow

Table 4 The 20 words selected only from chapter 8, and those only from chapter 9.

Chapter 8, 6 Areas

Rank of χ^2	Rank of Frequency	Word	Frequency	χ^2	Areas which shows the peak of χ^2
1	45	Express	37	7.18	6
2	31	Atom	53	6.69	5,6
3	47	Element	34	6.02	2
4	80	Molecular Formula	18	5.61	6
5	63	Particle	25	5.59	1
6	53	Sodium Chloride	31	5.31	1
7	55	Sodium Chlorate	28	5.03	1
8	79	Zinc	19	4.41	4
9	17	Molecule	94	4.33	2,5,6
10	57	Number	27	4.10	6
11	105	Formula of Chemical Reaction	12	3.73	6
12	49	Heat	33	3.30	1
13	104	Energy	12	3.20	4
14	95	Druggists' Scale	12	3.16	3
15	77	one	20	3.09	6
16	135	Flame	8	2.88	4
17	127	Size	9	2.79	5
18	165	Flame Reaction	6	2.75	2
19	111	Combine	11	2.60	5
20	102	Contain	12	2.49	2

Chapter 9, 4 Areas

Rank of χ^2	Rank of Frequency	Word	Frequency	χ^2	Areas which shows the peak of χ^2
1	68	Create	14	3.48	3
2	49	Electron	23	3.38	4
3	40	Small Lump	26	3.06	1
4	78	Heater Wire	13	3.01	3
5	86	Heat	12	2.78	3
6	87	Electric Power	12	2.78	3
7	63	Dry Cell	16	2.57	1
8	98	Calory	11	2.55	3
9	127	Water Calorimeter	8	1.85	3
10	88	Diode	12	1.76	4
11	53	♦	20	1.74	3
12	77	Time	13	1.71	3
13	71	Electricity	14	1.67	4
14	105	Electric Energy	9	1.62	3
15	41	Metalic Wire	25	1.56	2
16	76	Metal	13	1.53	4
17	67	Electric Registance	15	1.48	2
18	103	Electrode	10	1.47	4
19	104	Discharge	10	1.47	4
20	106	Terminal	9	1.45	1

Table 5 First four important words in individual chapter

(a) Chapter 8

area code	
1	Particle, Sodium Chloride, Sodium Chlorate, Heat
2	Element, Molecule, Flame, Flame Reaction
3	Druggists' Scale, Mass, Input, In and Out
4	Zinc, Energy, Burn, In and Out
5	Atom, Molecule, Size, Combine
6	Express, Atom, Molecular Formula, Molecule

(b) Chapter 9

area code	
7	Small Lump, Dry Cell, Terminal Circuit
8	Metalic Wire, Electric Resistance, Circuit, G
9	Create,Heater Wire, Heat, Electric Power
10	Electron, Diode, Electricity, Metal

When Chapter 8 and 9 are processed separately the result changes slightly as Table 4 and 5. Words "Material" in Chapter 8, "Current", "Voltage" and "Flow" in Chapter 9, which were important words in Table 2 and 3, fell off from Tables 4 and 5. This means that those words typify chapter contents, but not individual sections in a chapter.

4. Extraction of Important Words in Research Papers on Electrical Engineering

The same experiment is done for the research papers on electrical engineering. Japan Information Center of Science and Technology (JICST) publishes every month the abstracts in the field of science and technology in the form of printed books and magnetic tapes. We bought magnetic tapes in electrical engineering. The number of abstracts was 4,889, the total number of words was 120,050, and the number of different words was 33,764. Electrical engineering was subdivided into 19 areas (n=19) as shown in Table 6. The result is good by our common sense, but it is difficult to evaluate whether the result is satisfactory.

Table 6 The division of 'Current Bibliography on Science and Technology'

A. Electricity in General
1. Electricity in General 2. Electric Materials and Parts

B. Measurement Control
3. Measurement 4. Control

C. Electric Power Engineering
5. Electric Power 6. Power Apparatus 7. Power Applications

D. Electronic Engineering
8. Electronic Engineering in General 9. Electronic Parts
10. Electronic Circuit 11. Quantum Electronics
12. Applications of Electronic Techniques

E. Communication Engineering
13. Communication in General 14. Electric Wave Propagation, Antenna
15. Transmission Method and Devices, 16. Applications of Communication Engineering

F. Information Processing
17. Information Processing 18. Electronic Computer
19. Applications of Information Processing

5. Conclusion

The experiment was generally good. But there are a few specific problems originated from the characteristics of Japanese language, and several other general problems. We are testing several variations for the extraction algorithm of important words, and want to have much more satisfactory results by processing much more data.

References

(1) G. Salton : Recent Studies in Automatic Text Analysis and Document Retrieval, JACM, Vol.20, No. 2, April 1973

(2) M. Nagao, K. Ochiai, M. Mizutani: Automatic Extraction of Important Words by Using χ^2Test for the Information Retrieval of Japanese Documents, Annual Conference of IECEJ, No. 1451, July 1974