Modeling Content Identification from Document Images

Takehiro Nakayama Fuji Xerox Palo Alto Laboratory 3400 Hillview Avenue Palo Alto, CA 94304 USA nakayama@pal.xerox.com

Abstract

A new technique to locate content-representing words for a given document image using abstract representation of character shapes is described. A character shape code representation defined by the location of a character in a text line has been developed. Character shape code generation avoids the computational expense of conventional optical character recognition (OCR). Because character shape codes are an abstraction of standard character code (e.g., ASCII), the mapping is ambiguous. In this paper, the ambiguity is shown to be practically limited to an acceptable level. It is illustrated that: first, punctuation marks are clearly distinguished from the other characters; second, stop words are generally distinguishable from other words, because the permutations of character shape codes in function words are characteristically different from those in content words; and third, numerals and acronyms in capital letters are distinguishable from other words. With these classifications, potential content-representing words are identified, and an analysis of their distribution yields their rank. Consequently, introducing character shape codes makes it possible to inexpensively and robustly bridge the gap between electronic documents and hard-copy documents for the purpose of content identification.

1 Introduction

Documents are becoming increasingly available in machine-readable form. As they are stored automatically and transferred on networks, many natural language processing techniques that identify their content have been developed to assist users with information retrieval and document classification. Conventionally, stored records are identified by sets of keywords or phrases, known as index terms (Salton, 1991).

Although documents are increasingly being computer generated, they are still printed on paper for reading, dissemination, and markup. As it is believed that paper will remain a comfortable medium for reading and modification (O'Gorman and Kasturi, 1992), development of content identification techniques from a document image is still important. OCR is often used to convert a document image into machine-readable form, but processing performance is limited by the overhead of OCR (Mori *et al.*, 1992; Nagy, 1992; Rice *et al.*, 1993). Because of the inaccuracy and expense of OCR, we decided to avoid using it.

Instead, we have developed a method that first makes generalizations about images of characters, then performs gross classification of the isolated characters and agglomerates these character shape codes into spatially isolated (word shape) tokens (Nakayama and Spitz, 1993; Sibun and Spitz, this volume). Generating word shape tokens is inexpensive, fast, and robust. Word shape tokens are a potential alternative to character coded words when they are used for language determination and part-of-speech tagging (Nakayama and Spitz, 1993; Sibun and Spitz, this volume; Sibun and Spitz, 1993; Sibun and Spitz, this volume; Sibun and Farrar, 1994). In this paper, we describe an extension of our approach to content identification.

2 Word shape token generation from image

In this section, we introduce word shape tokens and their generation from document images.

First, we classify characters by determining the characteristics of the text line. We identify the positions of the baseline and the x-height as shown in figure 1 (Spitz, 1993).

Next, we count the number of connected components in each character cell and note the position of those connected components with respect to the baseline and x-height (Nakayama and Spitz, 1993; Sibun and Spitz, this volume). The basic character classes {A x i g j U '-, .: = !} and the members which constitute those classes are shown in Table 1. In this paper, they are represented in bold-face type (e.g., Aigxx). Note that a character shape code subset $\{-, . : !\}$ includes only punctuation marks. This is important for our cleaning process which will be described later.



Figure 1: Text line parameter positions

Character shape codes are grouped by word boundary into word shape tokens (see Sibun and Spitz, this volume). The correspondence between the scanned word image and the word shape token is one-to-one; that is, when a certain word shape token is selected, its original word image can be immediately located.

Recognizing word shape tokens from images is two or three orders of magnitude faster than conventional OCR (Spitz, 1994), and is robust for real-world documents which are sometimes degraded by poor printing and which sometimes use more than a single font.

character shape code	members		
Α	A-Z bdfhkltB 0-9 +#\$&()/<>[]@{ }		
X	acemnorsuvwxz		
i	iádaédeîoûù		
g	gpqyç		
j	j		
U	äëöüÖÜ		
'-,.:=!	·'-,.:;=!?		

Table 1: Shape class code membership

The use of only 13 character shape codes instead of approximately 100 standard character codes results in a one-to-many correspondence between word shape tokens and words.

Figure 2 shows how much character shape codes reduce word variation using the most frequent 10,000 English words (Carroll *et al.*, 1971) in order of frequency. A word is defined as a string of graphic characters bounded on the left and right by spaces. Words are distinguished by their graphic characters. For example, "apple", "Apple", and "apples" are three different words, while "will" (modal) and "will" (noun) are the same. For the purpose of comparing the character shape code representation with the standard character code representation, the x axis represents the number of frequent words, and the y axis represents the number of distinct words represented in both ASCII and character shape codes. The number of words in ASCII naturally corresponds to the number of original words one-toone. On the other hand, the number of words in character shape codes (the number of word shape tokens) is less than half of the number of original words. This gap is a constraint on the accuracy of our approach, but we show it is not a serious limitation in the following section.



Figure 2: ASCII and character shape code

3 Content identification

Text characterization is an important domain for natural language processing. Many published techniques utilize word frequencies of a text for information retrieval and text categorization (Jacobs, 1992; Cutting *et al.*, 1993).

We also characterize the content of the document image by finding words that seem to specify the topic of the document. Briefly, our strategy is to identify the frequently occurring word shape tokens.

In this section, we first describe a process of cleaning the input sequence which precedes the main procedures. Then, we illustrate how to collect the important tokens, introducing a stop list of common word shape tokens which is used to remove the tokens that are insufficiently specific to represent the content of the documents.

3.1 Cleaning input sequence

Given a sequence of word shape tokens, the system removes the specific character shape codes '-', ',', '.', ':', and '!' that do not contribute important linguistic information to the words to which they adhere, but that change the shape of the tokens. Otherwise, word shape would vary according to position and punctuation, which would interfere with token distribution analysis downstream. We ignore possible sentence initial word shape alteration by capitalization simply because it is almost impossible to presume the original shape. In this paper, capitalized words are counted differently from their uncapitalized counterparts.

Our cleaning process concatenates word shape tokens before and after the hyphen at the end of line. The process also deletes intended hyphens (e.g., AxxxA-xixAxA [broad-minded] --> AxxxAxixAxA). Eliminating hyphens reduces the variation of word shape tokens. We measured this effect using the aforementioned frequent 10,000 words. Forty-two words of 10,000 are hyphenated. In character shape code representation, 10,000 words map into 3,679 word shape tokens (figure 2). When hyphens are eliminated, the 10,000 words fall into 3,670 word shape tokens. This small reduction implies that eliminating hyphens does not practically affect the following process.

3.2 Introducing a word shape token stop list

After cleaning is done, the system analyzes word shape token distribution. Word shape tokens are counted on the hypothesis that frequent ones correspond to words that represent content; however, tokens that correspond to function words are also very frequent. One problem awaiting solution is that of developing a technique to separate these two classes.

In this paper, we define function words as the set of {prepositions, determiners, conjunctions, pronouns, modals, be and have surface forms}, and content words as the set of {nouns, verbs (excluding modals and be and have surface forms), adjectives}. Words that belong in both categories are defined as function words. We exclude adverbs from both, because they sometimes behave as function words and sometimes as content words. Words that can be adverbs but also can be either a function or a context word are not counted as adverbs.

In English, function words tend to be short whereas content words tend to be long. For the purpose of investigating characteristics of function and content words in character shape code representation, we compiled a lexicon of 71,372 distinct word shape token entries from an ASCII-represented lexicon of 245,085 word entries which was provided by Xerox PARC and was modified in our laboratory. 254 word shape token entries of the lexicon correspond to 515 function words, 63,356 entries correspond to 226,648 content words, and 209 entries correspond to both function and content words. Finally, 8,921 word shape token entries correspond to 17,922 adverbs. Figure 3 shows the distribution of word shape token length. Frequency of occurrence of word shape tokens was not taken into account; that is, we simply counted the length of each entry and computed the population ratio. The distribution of content words is apparently different from that of function words. In the figure, we also record the distribution of word shape tokens corresponding to the 100 most frequent words (75 function words, 16 content words, and 9 adverbs) from the source (Carroll et al., 1971). It illustrates that very common words are short.



Figure 3: Word shape token length distribution

A stop list of the most common function word shape tokens was constructed so that they could be removed from sequences of word shape tokens. It is important to select the right word shape tokens for this list, which must selectively remove more function words than content words. In general, the larger the list, the more it removes both function and content words. Thinking back to our goal of finding frequent content words, we don't need to try to remove all function words. We need only to remove the function words that are usually more frequent than content-representing frequent words in the text on the assumption that the frequency of individual function words is almost independent of topic. Infrequent function words that remain after using the word shape token stop list are distinguishable from frequent content words by comparing their frequencies.

We generated a word shape token stop list using Carroll's list of frequent function words. We selected several sets of the most frequent function words, by limiting the minimum frequency of words in the set to 1%, 0.5%, 0.1%, 0.09%, 0.08%, ..., 0%, then converted them into word shape tokens. We tested these word shape tokens on the aforementioned lexicon to count the number of matching entries. Table 2 gives part of the results, where Freq.FW stands for frequencies of the selected function words, # FW for the number of them, # stop-tokens for the number of word shape tokens derived from them, FW.Match for a ratio of the number of matching function words to the total number of function words in the lexicon (515), and CW.Match for a ratio of the number of matching content words to the total number of content words (226,648). A word shape token stop list, for instance, from function words whose frequencies are more than 0.5% removes 0.4% of content words and 18% of function words from the lexicon; a word shape token stop list from function words with frequencies more than 0.01% removes 4.2% of content and 56% of function words; and a word shape token stop list from all function words in the lexicon removes 9.5% of content words.

Table 2: Application	of word	shape token	stop list to
lexicon			

Freq.FW (%)	# FW	# stop- tokens	FW. Match (%)	CW. Match (%)
>1	7	6	6.2	0.1
> 0.5	19	15	18	0.4
> 0.1	77	44	39	1.8
> 0.09	81	44	39	1.8
> 0.07	87	47	41	2.4
> 0.05	101	57	44	2.9
> 0.03	115	68	49	3.7
> 0.01	153	95	56	4.2
>0	515	254	100	9.5

Function words (frequency > 0.05%)

the of and a to in is you that it he for was on are as with his they at be this from I have or by one had but what all were when we there can an your which their if will each about up out them she many some so these would other into has her like him could no than been its who now my over down only may after where most through before our me any same around another must because such off every between should under us along while might next below something both few those

Word shape token stop list frem above words

AAx xA xxA x Ax ix gxx AAxA iA Axx xxx xx xiAA Aix AAxg AAix Axxx A Ag AxA xAxA xAA xxxx xAxx AAxxx gxxx xAixA AAxix xxxA xAxxA xg AAxx xAx xxxg xxxAA xAAxx ixAx AiAx iAx xxAg xxg xAxxx AAxxxgA AxAxxx xxxxAA xxAAxx Axxxxxx xxxxg AxAxxxx xAxxAA xxAAxx xAxxg xAiAx xigAA AxAxx xxxxAAixg AxAA

Figure 4: Selected function words and word shape token stop list

We also tested these word shape token stop lists on ASCII encoded documents, and discovered that good results are obtained with the lists derived from function words with frequencies of more than 0.05%. This list identifies all words that occur more than 5 times per 10,000 in the document. Figure 4 shows the selected function words and the corresponding word shape token stop list. The number of stop tokens is 57 for 101 function words. Table 2 shows that the list removes 2.9% of content words and 44% of function words from the lexicon.

3.3 Augmentation of the word shape token stop list

In our character classification, all numeric characters are represented by the character shape code A (Table 1). Therefore, after cleaning is done, all numerals in a text fall into word shape tokens A^* , where * means zero or more repetitions of A. This sometimes makes the frequency of A^* unreasonably high though numerals are often of little importance in content identification.

 A^* matches all numerals, but since it matches few content words except for acronyms in capital letters, we decided to add A^* to the word shape token stop list.

sample	$\begin{array}{c} \text{CW.1} \\ \rightarrow \text{CW.2} \end{array}$	CW.R (%)	FW.1 → FW.2	FW.R (%)
doc.1	347 → 321	7.5	74 → 18	76
doc.2	246 → 221	10	63 → 7	89
doc.3	245 → 225	8.2	61 → 10	85
doc.4	292 → 272	6.8	61 → 7	89
doc.5	279 → 265	5.0	71 → 16	78
doc.6	255 → 236	7.5	56 → 12	79
doc.7	177 → 164	7.3	53 → 14	74
doc.8	253 → 231	8.7	71 → 17	76
doc.9	227 → 214	5.7	64 → 11	83
doc.10	239 → 218	8.8	63 → 14	78
doc.11	233 → 212	9.0	62 → 10	84
doc.12	294 → 265	9.9	58 → 12	79
doc.13	233 - → 212	9.0	57 → 12	79
doc.14	271 → 248	8.5	59 → 13	78
doc.15	130 → 115	12	42 → 5	88
doc.16	1582 → 1513	4.4	150→ 45	70
doc.17	453 → 409	9.7	99 → 17	83
doc.18	292 -→ 249	15	75 → 8	89
doc.19	1189 -→ 1046	12	157→ 35	78
doc.20	309 → 286	7.4	73 → 6	92

Table 3: Testing the word shape token stop list on sample documents

3.4 Testing the word shape token stop list

Our word shape token stop list was tested on 20 ASCII encoded sample documents, ranged in length from 571 to 13,756 words, from a variety of sources including business reports, travel guides, advertisements, technical reports, and private letters. First, we generated word shape tokens, and cleaned as described earlier. Next, we removed tokens that were on the word shape token stop list. Table 3 shows the number of content and function words which the documents consist of before and after using the list. In the table, CW.1 and CW.2 stand for the number of distinct content words in the original document and the number after using the word shape token stop list, respectively. CW.R stands for a ratio of (CW.1 - CW.2) to CW.1. Similarly, FW.1 and FW.2 stand for the number of distinct function words before and after using the list, and FW.R is a ratio of (FW.1 -

FW.2) to FW.1. FW.R is much larger than CW.R in all sample documents, which shows the good performance of the word shape token stop list. We should note that the values of CW.R are larger than the 2.9% that we get from testing the list on the lexicon. This is because the lexicon includes many uncommon words and these tend to be longer than the function words selected to make the word shape token stop list. This implies that our list removes more ordinary content words than uncommon ones. We believe that removing such words affects content identification little since ordinary content words in many topics usually don't

(count)	
67 AAx	{the, The}
55 Ax	{to, be, Fr, In, As, On, An, (e}
35 AAAA	{1988, 1989, 2000, 1990, 1987, +7%}, +5%}, +28%, +27%, +18%, +14%}
33 xA	{of, at}
32 xxA	{and, out, not, end, act}
31 ix	{in, is}
29 Axx	{for, due, ten, low, For, Enz}
28 xxxx	{some, over, were, more, same, rose, ease}
23 AA	{6%, 5%, At, 9%, 7%, 4%, 3%, 1%, 8%, 2%, 11, 10, +6}
22 AxiAAixg	{building, Building}
22 A	{4, 9, 8, 6, 3, 2, 1, 0, R, A, 5}
19 xxxA	{work, real, cost, such, much, most}
18 AAA	{90%, 83%, 847, 80%, 5%}, 49%, 29%, 27%, 26%, 23%, 21%, 19%, 175, 14%, 13%}
16 x	{a, s}
16 gxxx	{year, pace, grew}
16 Ag	{by, By}
14 xxx	{was, are, new, can, saw, own, one}
12 xxxxAxxxAixx	{construction}
10 xxgxxAxA	{expanded, expected, reported}
9 xx	{on, as, or}

After using the word shape token stop list: word shape token ranking and corresponding words

22 AxiAAixg	{building, Building}	6 Aixxx	{firms, Since}
12 xxxxAxxxAixx	{construction}	6 AixxA	{first, fixed}
10 xxgxxAxA	{expanded, expected, reported}	5 ixxxxxxx	{increase}
8 xxxAxx	{sector, number}	5 AxxxxA	{demand, traced, lowest}
7 xxgixxxxixg	{engineering}	5 Axxxixg	{housing}
7 xxAxxx	{volume, orders, return}	5 AiAAixx	{billion}
7 ixAxxAxixA	{industrial}	4 xxxAxxxAx	{contracts}
7 gxxxAA	{growth}	4 xxAx	{rate}
7 gxixxx	{prices}	4 ixAxxixx	{interior}
7 AxAxAxx	{October}	4 gxxxxAixg	{preceding}
6 xxxxixgx	{earnings}	4 gxixA	{point}
6 AxxAx	{trade}	4 AxxxxAxx	{branches}
6 AxxAAxg	{backlog}	4 Axixx	{Swiss}

Figure 5: Most frequent word shape tokens and corresponding words

specify the content of the document well (Salton *et al.*, 1975). Likewise, the values of FW.R are larger than 44% for the same reason.

After using the word shape token stop list, we counted the remaining tokens to obtain content-representing words in their frequency order. All samples successfully indicated their content by appropriate words.

Data for a sample document reporting the growth of the building industry in Switzerland in 1988 and its outlook for 1989, consisting of 1013 words, are shown in figure 5. It shows top frequent word shape token ranking of the original document and the new ranking after using the word shape token stop list. The number of removed tokens was 544. Most of them represented common words and numerals. The top ranking after using the word shape token stop list consists of content words and represents the content of the document much better than the ranking before using it.

Figure 5 also suggests that we can inexpensively locate key words by performing OCR on the few frequent word shape tokens.

4 Conclusions and further directions

Generating word shape tokens from images is inexpensive and robust for real-world documents. Word shape tokens do not carry alphabetical information, but they are potentially usable for content identification by locating content-representing word images. Our method uses a word shape token stop list and analyzes the distribution of tokens. This technique depends on the observation that, in English, the characteristics of word shape differ between function and content words, and between frequent and infrequent words.

We expect to be able to extend the technique to many other European languages that have similar characteristics. For example, German function words tend to be shorter than nouns, which are always capitalized. In addition, by drawing on our language determination technique, which uses the same word shape tokens (Nakayama and Spitz, 1993; Sibun and Spitz, this volume), we could enhance the technique described here for multilingual sources.

Other future work involves examining automatic document categorization in which an input document image is assigned to some pre-existing subject category (Cavnar and Trenkle, 1994). With reliable training data, we feel we can identify the configuration of word shape tokens across subjects. Using a statistical method to compute the distance between input and configurations of categories would be a good approach. This might be useful for document sorting service for fax machines.

Acknowledgments

The author gratefully acknowledges helpful suggestions by Larry Spitz and Penni Sibun.

References

- John B. Carroll, Peter Davies, and Barry Richman, The American Heritage word frequency book, Boston, Houghton-Mifflin, 1971.
- William B. Cavnar and John M. Trenkle, N-Gram-Based Text Categorization, Proceedings of the Third Annual Symposium on Document Analysis and Information Retrieval, Las Vegas, U.S.A., 1994.
- Douglass R. Cutting, David R. Karger, and Jan O. Pedersen, Constant Interaction-Time Scatter/Gather Browsing of Very Large Document Collections, Proceedings of the 16th Annual International ACM SIGIR Conference, Pittsburgh, U.S.A., 1993.
- Paul S. Jacob, Joining Statistics with NLP for Text Categorization, Proceedings of the Third Conference on Applied Natural Language Processing, Trento, Italy, 1992.
- Shunji Mori, Ching Y. Suen, and Kazuhiko Yamamoto, Historical Review of OCR Research and Development, Proceedings of IEEE, Vol. 80, No. 7, 1992.
- George Nagy, At the Frontiers of OCR, Proceedings of IEEE, Vol. 80, No. 7, 1992.
- Takehiro Nakayama and A. Lawrence Spitz, European Language Determination from Image, Proceedings of the Second International Conference on Document Analysis and Recognition, Tsukuba Science City, Japan, 1993.
- Lawrence O'Gorman and Rangachar Kasturi, Document Image Analysis Systems, Computer July 1992 Vol. 25.
- Stephen V. Rice, Junichi Kanai and Thomas A. Nartker, An Evaluation of OCR Accuracy, Information Science Research Institute 1993 Annual Research Report, University of Nevada, Las Vegas.
- Gerard Salton, Developments in Automatic Text Retrieval, Science Vol. 253, No. 5023, Aug. 30, 1991.
- Gerard Salton, A. Wong, and C. S. Young, A Vector Space Model for Automatic Indexing, Communications of the ACM November 1975 Vol. 18.
- Penelope Sibun and David S. Farrar, Content Characterization Using Word Shape Tokens, Proceedings of the 15th International Conference on Computational Linguistics, Kyoto, Japan, 1994.
- Penelope Sibun and A. Lawrence Spitz, Language Determination: Natural Language Processing from Scanned Document Images, this volume.
- A. Lawrence Spitz, Generalized Line Word and Character Finding, Proceedings of the International Conference on Image Analysis and Processing, Bari, Italy, 1993.
- A. Lawrence Spitz, Using Character Shape Codes for Word Spotting in Document Images, Proceedings of the Third International Workshop on Syntactic and Structural Pattern Recognition, Haifa, Israel, 1994.