A Web Search Result Classification System Based on the Degree of Suitability for Specialists

Masanobu Tsuruta Yoshiyuki Umemura Hiroyuki Sakai Shigeru Masuyama
Toyohashi Univercity of Technology
1-1 Hibarigaoka, Tenpaku-cho, Toyohashi-shi, Aichi 441-8580, Japan
{tsuruta,umemura,sakai,masuyama}@smlab.tutkie.tut.ac.jp

Abstract

We propose a search result classification system based on the degree of suitability for specialists. Our system classifies documents based on whether a document is "for specialists" or "not for specialists". This classification is done by ranking each document using the density of keywords and by a threshold. For evaluation of our system, we participated in NTCIR4-WEB Topical classification task. Only a part of "evaluation based on distribution of relevant documents" has executed at this time. In this evaluation, our system exhibits good performance for some data set. We also independently conducted experiments for evaluation based on classification errors. Experimental results show that the precision of classification of our system attains value close to that of human evaluators' average.

Keywords: *search result classification system, density of key words, suitability for specialists.*

1 Introduction

The information retrieval and document classification technologies [1] are indispensable for efficient use of various documents on the Web. Most existing classification systems were designed to group documents according to categories[5]. Such systems do not take document suitability for specialists into consideration, therefore, both types of documents, for specialists, and those not for specialists, respectively, are included in their classification results. Then, a user needs to select documents matched to one's knowledge.

To cope with this problem, we propose a Web-search results classification system which uses the measure of document suitability for specialists. The conventional classification systems use the content similarity between some documents, but our system does not use it. Therefore, our system cannot group documents according to categories.

2 System description

2.1 Definition of "suitability for specialists"

First, we define "suitability for specialists." In our system, a document's suitability for specialists denotes "when he reads the document, how much does the reader need special knowledge relevant to the most suitable field which the document belongs to?"

2.2 Features of documents for specialists, and not for specialists

As a preliminary study, we classified some documents according to suitability for specialists, and we examined features of documents for specialists, and those not for specialists, respectively. Most documents for specialists have the following features.

- Documents for specialists were written in argumentative, difficult words.
- Abbreviations, acronyms, jargons and technical terms, as we say "keywords" frequently appears in documents for specialists.
- They contain much "information."

In contrast, most documents not for specialists have the following features.

- Documents not for specialists were written in easy expression.
- General phrases are used.
- They contain little "information" comparing with those for specialists.

2.3 Keywords and their density

Keywords frequently appear in documents for specialists. On the other hand, keywords do not often appear in documents not for specialists. From these observations, we decided to classify documents based

on suitability for specialists, by using density of keywords appearance in a document for the measure how the one is suitable for specialists.

2.3.1 Definition of keywords

We formally define a "keyword" and a "compound noun". A keyword is a *Katakana* noun or a compound noun which often appears in documents of Web-search result. A compound noun is a combination of at least 2 nouns. Our system uses the longest compound noun in all compound nouns made by a sequence of nouns. We will explain the reason in section 2.4.2.

2.4 Method of classification based on suitability for specialists

Step 1: Select compound noun $t_i, i = 1, 2, ..., n$, contained in document set S of search results.

Step 2: Compute weight $W(t_i, S)$ of compound noun t_i in document set S.

$$W(t_i, S) = (A + \frac{Tf(t_i, S)}{\max_i Tf(t_i, S)}) \times \log \frac{|N|}{df(t_i, N)} \times (B + \frac{En(t_i, S)}{\max_i En(t_i, S)}), \quad (1)$$

where
A, B: constants[1]
$Tf(t_i, S)$: Frequency of appearance of compound noun t_i in document set S.

$$Tf(t_i, S) = \sum_{s \in S} tf(t_i, s), \quad (2)$$

where
$tf(t_i, s)$: Appearance frequency of compound noun t_i in document s.
$En(t_i, S)$: Entropy based on appearance probability of compound noun t_i in document set S

$$En(t_i, S) = -\sum_{s \in S} P(t_i, s) \log_2(P(t_i, s)) \quad (3)$$

$$P(t_i, s) = \frac{tf(t_i, s)}{Tf(t_i, S)} \quad (4)$$

$df(t_i, N)$: Frequency of documents which contain compound noun t_i in entire document set N.

Step 3: Adopt m best compound nouns with respect to ranking $W(t_i, s)$ in W for keywords.

[1] we define $A = 0.4, B = 0.4$ by trial and error.

Step 4: Compute density of keywords $Den(s)$ for each document[4].

$$Den(s) = \frac{\sum_{t \in KS(s)} W(t, S)}{d(s)} \quad (5)$$

$$d(s) = \frac{\sqrt{\sum_{k=2}^{|KS(s)|} (dist_k)^2}}{|KS(s)| - 1} \quad (6)$$

$KS(s)$: Set of keywords appeared in s.
$dist_k$: Distance between t_k and t_{k-1} (the number of words.)

Step 5: Put documents in the order of $Den(s)$.

Step 6: Set the threshold T, and classify the document s as "for specialists" if $Den(s) > T$, otherwise, classify s as "not for specialists." □

2.4.1 Weighting function

The formula 1 is a weighting function which assigns heavy weights to the compound nouns that often appear only in the Web-search results.

The first term assigns the normalized appearance frequency of compound nouns in the entire documents included in the Web-search result. If a compound noun has the highest frequency of appearance, this term gives it the largest value.

The second term assigns inverse document frequency. If a compound noun has the highest probability of appearing at least one in each document, this term gives it the smallest value.

The third term assigns the normalized entropy based on the probability of appearance frequency of a compound noun. If a compound noun appears only in a document, this term gives it the smallest value. If the weighting function is the product of the first and the second terms without the third term, the function assigns not heavy weights to such compound nouns that appear only in a document, and appear frequently in the document. Such compound nouns are not appropriate as keywords because keywords in the sense of this paper are the delegates of not a document but those of a document set. Therefore, lighter weights are assigned to such compound nouns by the third term. Keywords are assigned heavy weights by the combination of these three terms.

2.4.2 Density of Keywords

The formulas 5 and 6 assign the density of keywords $Den(s)$ [4][3]. The formula 6 assigns the squared average distance between keywords appearing in a document. In figure 1, the variants of formula 6 have the following values: $|KS(S_i)| = 3$, $dist_2 = 2$ and

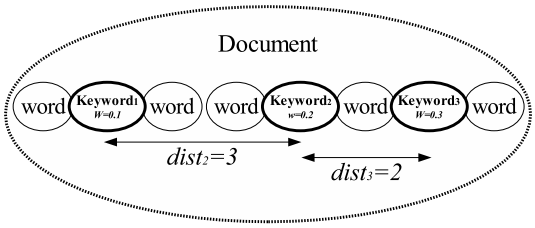

Figure 1. Density of Keywords

$dist_3 = 1$. If the squared average distance of a document is short, keywords appears densely in the document.

On the Web, some documents consist of multiple articles. Moreover, such articles in a document may have different subjects. When one assigns the densities of keywords of such documents, the document that includes important and technical contents only in a part of the document might be classified as not for specialists, if we use the simple distance of keywords. We try to classify such documents appropriately by using the densities of keywords computed by the squared average distance. In addition, the reason why our system always use the longest compound noun is as follows: The following problem arises, when computing densities of keywords.

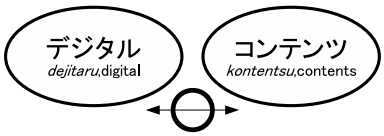

Figure 2. Distance between two nouns, "デジタル (*dejitaru*,digital)" and "コンテンツ (*kontentsu*,contents)"

Figure 3. Distances between a compound noun "デジタルコンテンツ (*dejitaru kontentsu*,digital contents)" and a vicinity keyword

If we do not use any compound noun as a keyword, we can easily assign the density of keywords because we can estimate the distance between a keyword and the next keyword as shown in figure 2. If we always use the longest compound noun as a keyword,

Figure 4. Distance between compound noun "デジタルコンテンツ (*dejitaru kontentsu*,digital contents)" and a noun "コンテンツ (*kontentsu*,contents)" which is a part of "デジタルコンテンツ (*dejitaru kontentsu*,digital contents)"

as shown in figure 3, no problem arises. However, if we use both compound and nouns as keywords, a problem arises. Consider the case when "デジタルコンテンツ (*dejitaru kontentsu*,digital contents)," the compound noun and "コンテンツ (*kontentsu*,contents)," a noun is adopted as a keyword, as illustrated in figure 4. Then, it is impossible for us to assign the density of keywords in such a case. We want to use any noun(in particular, a proper noun) as a keyword, but, at this point, we do not know how to cope with such a case. Therefore, we define our system to use only the longest compound noun.

2.4.3 Threshold

To classify documents by our system, we must define a threshold. Thus, we conducted preliminary experiments for determination of the threshold.

Preliminary experiments for determination of the threshold Three engineering students classify up to 200th documents of a Web-search result, and from those documents, we make a correct-data for a preliminary experiments by the majority decision. Therefore, we define a threshold from the proportion of the number of documents classified as "for specialists" and as "not for specialists" in the correct-data.

Result of the preliminary experiments In the correct-data, 91 documents are classified as "for specialists", 109 documents are classified as "not for specialists." By the result, our system has defined the median as the threshold.

3 Evaluation of our system

Two kinds of evaluations were conducted. Our participant-ID is smlab-01.

3.1 Evaluation based on distribution of relevant documents

The organizer announced two relevance judgement data and the results of evaluation based on distribution of relevant documents.

There are two relevance judgement datas, “META-02.relax.res” and “META-02.rigid.res”. The “rigid” one is more strict about relevance judgement of documents than the “relax” one. We show the results on this evaluation.

Table 1 shows the result on “trec_eval” [2] with the relax relevance judgement data, table 2 shows the result on “trec_eval” with the rigid relevance judgement data. Table 3 shows the result on the DCG measures with the relax data, and Table 4 shows the rigid data.

The DCG measures are computed by the following formulas.

$$cg(i) = \begin{cases} g(1), & \text{if } i = 1 \\ cg(i-1) + g(i), & \text{otherwise} \end{cases}$$

$$dcg(i) = \begin{cases} g(1), & \text{if } i = 1 \\ dcg(i-1) + \frac{g(i)}{\log_2(i+1)}, & \text{otherwise} \end{cases}$$

$$mdcg_1(i) = \begin{cases} g(1), & \text{if } i = 1 \\ mdcg_1(i-1) & \\ +\frac{g(i)}{\log_2(i+1)\times\log_2 k(j+1)}, & \text{otherwise} \end{cases}$$

$$mdcg_2(i) = \begin{cases} g(1), & \text{if } i = 1 \\ mdcg_2(i-1) + \frac{g(i)}{\log_2(j+1)}, & \text{otherwise} \end{cases}$$

where, $k = 1$,
i: position of the document in the order (1-20),
j: position of the class in the order,

$$g(i) = \begin{cases} 1, & \text{if the document is highly,} \\ & \text{fairly or partially relevant.} \\ 0, & \text{otherwise.} \end{cases}$$

3.1.1 Evaluations method based on distribution of relevant documents

The organizer announced that the evaluation was conducted according to the following steps.

Step 1. Rank the classes of the classification result in the order, from the class which includes relevant documents the most. In the rest of this paper, we call the result on this step data 1 with respect to the rank assigned in Step 1.

Step 2. Adopt up to 20th documents from data 1. If the method is a non-exclusive method, and if the same documents appear in multiple classes, then retain the name of the former documents in data 1, but mark “d:” for the head of the id of the other documents. The documents marked does not seem relevant on “trec_eval.” In the rest of this paper, we call the result on this step data 2.

Step 3. Run “trec_eval with datas 1 and 2.

Step 4. Compute the DCG measures with datas 1 and 2.

[2] ftp://cs.cornell.edu/pub/smart/trec_eval.v3beta.shar

3.1.2 Results on the evaluation based on the distribution of relevant documents

Evaluation results of our system exhibit good scores of precision, recall and F value when there are a number of relevant documents in the target data set. In addition, evaluation results of our system exhibit good score of DCG measure in such a target data set.

3.2 Evaluation based on classification errors

Organizer has not yet provided the evaluation result based on classification errors. Therefore, we independently conducted this evaluation.

We cannot evaluate the performance for extracting the structure of classes of our system, because our system classifies documents for two particular classes. Thus, we did not evaluate the performance of the classification to the structure of classes.

3.2.1 Evaluation method based on classification errors

On this independent evaluation, our system classified one of target data set which used at NTCIR4 WEB task D dry-run, topic id 0028, a search result for “著作権 (*chosakuken*,copyright), デジタルコンテンツ (*dejitaru kontentsu*,digital contents), ネットワーク (*nettowaaku*)” (copyright, digital contents, network) that retrieved from the 100-gigabyte data set ‘NW100G-01’. In this study, we use up to 200th documents in the target data set.

We use Mecab [3] as a morphological analyser. Then, we use up to 100,000th compound nouns with respect to ranking $W(t_i, S)$ in W order for keywords. First, three engineering students classify documents of target data set, “for specialists” and “not for specialists.” From those majority decision, we make a correct-data. Then, we compare correct-data with the result on classification by our system, and compute scores of precision and recall.

3.2.2 Result on the evaluation based on classification errors

Both the recall and precision are 0.68. Human evaluators’ recalls were 0.94, 0.69 and 0.83, respectively, where, a human evaluator’s recall R_{HE} is computed by the following formula.

$$R_{HE} = \frac{|S_{s,e}|}{|S_{s,m}|}$$

where, $S_{s,e}$ is the document set classified as “for specialists” by a human evaluator,
$S_{s,m}$ is the document set classified as “for specialists” by the answer-data.

[3] http://cl.aist-nara.ac.jp/%7Etaku-ku/software/mecab/

Table 1. Result on 'trec_eval' with META-02.relax.res

		#RelDocRet		AvePrec		Prec@20		Recall@20		Fvalue@20	
query	#AllRelDoc	Smlab-01	avg	Smlab-01	avg	Smlab-01	avg	Smlab-01	avg	Smlab-01	avg
1	10	0	5.63	0	0.32	0	0.28	0	0.56	0	0.38
3	24	8	8.25	0.13	0.21	0.4	0.41	0.33	0.34	0.36	0.38
4	3	0	2	0	0.14	0	0.1	0	0.67	0	0.17
6	58	19	16.88	0.28	0.26	0.95	0.84	0.33	0.29	0.49	0.43
8	155	19	17.88	0.12	0.11	0.95	0.89	0.12	0.12	0.22	0.2
19	5	2	3.13	0.03	0.37	0.1	0.16	0.4	0.63	0.16	0.25
21	4	1	2.5	0.02	0.28	0.05	0.13	0.25	0.63	0.08	0.21
22	99	9	14.25	0.07	0.12	0.45	0.71	0.09	0.14	0.15	0.24
23	4	0	1.88	0	0.14	0	0.09	0	0.47	0	0.16
29	29	1	9.63	0	0.17	0.05	0.48	0.03	0.33	0.04	0.39
45	15	3	5.38	0.03	0.24	0.15	0.27	0.2	0.36	0.17	0.31
avg	36.91	5.64	7.94	0.06	0.21	0.28	0.4	0.16	0.41	0.15	0.28

Table 2. Result on 'trec_eval' with META-02.rigid.res

		#RelDocRet		AvePrec		Prec@20		Recall@20		Fvalue@20	
query	#AllRelDoc	Smlab-01	avg	Smlab-01	avg	Smlab-01	avg	Smlab-01	avg	Smlab-01	avg
1	10	0	5.63	0	0.32	0	0.28	0	0.56	0	0.38
3	15	6	5	0.13	0.13	0.3	0.25	0.4	0.33	0.34	0.29
4	3	0	2	0	0.14	0	0.1	0	0.67	0	0.17
6	58	19	16.88	0.28	0.26	0.95	0.84	0.33	0.29	0.49	0.43
8	155	19	17.88	0.12	0.11	0.95	0.89	0.12	0.12	0.22	0.2
19	5	2	3.25	0.03	0.36	0.1	0.16	0.4	0.65	0.16	0.26
21	3	1	2.25	0.03	0.33	0.05	0.11	0.33	0.75	0.09	0.2
22	71	6	11.75	0.03	0.12	0.3	0.59	0.08	0.17	0.13	0.26
23	1	0	1.38	0	0.55	0	0.07	0	1.38	0	0.13
29	11	0	6	0	0.22	0	0.3	0	0.55	0	0.39
45	11	2	6.13	0.02	0.3	0.1	0.31	0.18	0.56	0.13	0.4
avg	31.18	5	7.1	0.06	0.26	0.25	0.36	0.17	0.55	0.14	0.28

Observations for documents relevant to their class are as follows:

- Online glossaries and the bibliographies were classified as "for specialists."
- Some short supporting documentations were classified as "not for specialists."

Observations for documents irrelevant to their class are as follows:

- Individual journals and some message board not suitable for specialists, were classified as "for specialists."
- The documents mainly consisting of graphics, suitable for specialists, were classified as "not for specialists."

3.2.3 Discussion on the evaluation based on classification errors

Because the average of human evaluator's recall was 0.82, the result on our system was almost good. We made sure that the density of keywords were available for the measure of suitability for specialists.

Individual journals and some message boards not suitable for specialists, were classified "suitable for specialists." In such a "neighborhood", there are sometimes many abbreviations, acronyms, jargons that used only in that neighborhood. The keyword weighting function of our system does not assign the heavy weight of keyword which appears in only one document and appears frequently in the document. However, our system could not work on occasions that the target data set has some documents e.g., the log files of the same message board.

In addition, such a jargon often appeared, in paticular, articles which has any jargon. The density of keywords measured by squared average distance became high in such occasions.

There were documents whose density of keywords are zero. These documents did not have more than 2 keywords because these documents were too short. Therefore, the density of keywords of the documents were not computed.

4 Conclusion

We proposed a Web-search result classification system based on suitability for specialists.

Our system is still a prototype and we are planning to conduct the following functional improvements in our system:

Firstly, we will permit a user to change the proportions of classes. The proportions of classes in the target data set, will change because of the query word.

Table 3. Result on DCG measures with META-02.relax.res

	cg@20		dcg@20		$mdcg_1$@20		$mdcg_2$@20	
query	Smlab-01	avg	Smlab-01	avg	Smlab-01	avg	Smlab-01	avg
1	0	5.13	0	2.39	0	2.17	0	4.42
3	8	8	2.41	3.02	2.41	2.86	8	7.35
4	0	2	0	0.71	0	0.68	0	1.89
6	19	16.88	6.04	6.05	6.04	6.04	19	16.83
8	19	17.88	6.77	6.28	6.77	6.23	19	17.69
19	2	2.63	0.48	1.33	0.48	1.14	2	2.03
21	1	2.13	0.26	1.21	0.26	1.15	1	1.91
22	9	14.25	3.96	5	3.96	4.99	9	14.2
23	0	1.88	0	1.08	0	0.96	0	1.51
29	1	9.63	0.26	3.23	0.26	3.08	1	9.01
45	3	5.38	0.8	2.59	0.8	2.53	3	5.16
avg	5.64	7.8	1.91	2.99	1.91	2.9	5.64	7.45

Table 4. Result on DCG measures with META-02.rigid.res

	cg@20		dcg@20		$mdcg_1$@20		$mdcg_2$@20	
query	Smlab-01	avg	Smlab-01	avg	Smlab-01	avg	Smlab-01	avg
1	0	5.13	0	2.39	0	2.17	0	4.42
3	6	5	1.87	1.83	1.87	1.73	6	4.6
4	0	2	0	0.71	0	0.68	0	1.89
6	19	16.88	6.04	6.05	6.04	6.04	19	16.83
8	19	17.88	6.77	6.28	6.77	6.23	19	17.69
19	2	2.75	0.48	1.39	0.48	1.18	2	2.13
21	1	2.25	0.26	1.24	0.26	1.05	1	1.69
22	6	11.75	1.83	3.65	1.83	3.63	6	11.66
23	0	0.88	0	0.81	0	0.81	0	0.88
29	0	6	0	2.2	0	2.1	0	5.61
45	2	6.13	0.57	2.36	0.57	2.1	2	5.1
avg	5	6.97	1.62	2.63	1.62	2.52	5	6.59

By doing this, our system will be more useful. With a combination using writing style and densities of keywords, the precision will become better.

Secondly, we will adopt good keywords from all nouns except compound nouns to avoid the problem of zero-density of keywords which is caused by applying only compound nouns for keywords.

Finally, not using the longest compound noun, but finding the most effective statistical delimiting method, we will be able to make our system to compute the better density of keywords. Therefore, additional study about the way of delimiting compound nouns is desirable.

Acknowledgement

This work was supported in part by The 21st Century COE Program “Intelligent Human Sensing”. from the Ministry of Education, Culture, Sports,Science and Technology and The Grant-in-Aid ((C)(2)1368044) from the Japan Society for the Promotion of Science of Japan.

References

[1] R. Baeza-Yates and B. Ribeiro-Neto. *Modern Information Retrieval*. Addison Wesley, 1999.

[2] K. Eguchi. Readme.ntcweb4-d-v01.txt, 2004.

[3] C. Kwok, O. Etzioni, and D. Weld. Scaling quastion answering to the web. *Proc. 10th ACM-WWW*, 2001.

[4] H. P. Luhn. The automatic creation of literature abstracts. *Advances in Automatic Text Summarization*, 1999 (originally presented at IRE National Convention, 1958).

[5] C. D. Manning and H. Schütze. *Foundations of Statistical Natural Language Processing*. The MIT Press, 1999.