Feature Extraction Using Restricted Bootstrapping

廣川, 佐千男
九州大学情報基板研究開発センター

出版情報：2012 IEEE/ACIS 11th International Conference on Computer and Information Science (ICIS), pp.283-288, 2012-05. IEEE
バージョン：
権利関係：

Feature Extraction using Restricted Bootstrapping

Sachio Hirokawa
Research Institute for Information Technology
Kyushu Universy
Japan
hirokawa@cc.kyushu-u.ac.jp

***Abstract*—The bootstrapping method is known as an application of the Page-rank technique for documents and words. The technique calculates the score of the words by mutually propagating the score of the words and the documents. However, sometimes the result is far away from the initial query word. The problem is known as "topic drift". This paper proposes to restrict the words to be to the top t words in the process of bootstrapping. The method is simpler than the technique known so far. The method is applied for the real bankruptcy information documents to extract the bankruptcy causes strongly related to the query. It is confirmed that the method prevents the topic drift.**

***Keywords*-topic drift, feature word, bankruptcy information, text mining**

I. Introduction

Feature extraction of documents is one of the key issues in text mining. Characteristic keywords are crucial to understand or to summarize the search result and can be used for query expansion. The bootstrapping method decomposes the process of feature extraction into the document retrieval process and the keyword extraction and combines them as an iterated process. The score of a keyword is propagated to the next stage where the score of a document is evaluated based on the word score. After repetition of sufficient number of times, the importance of a word can be evaluated automatically, since the dignity of a word and the dignity of a document are converged on a fixed value.

The "topic drift" is the problem known to the method, where the keywords obtained tend to be too general compared to the original query. To prevent the topic drift, the conventional researches improve the evaluation method of words and documents dignity, or consider heuristics to restrict the occurrence pattern of the words. However, they loose the simplicity and generality of the original bootstrapping method.

Propagation of the word score and the document score is a key idea of bootstrapping. However, the propagation causes the accumulation of scores, and many a little makes a mickle. In this paper, we propose to cut the score propagation by using ranking instead of scores. When we retrieve the documents in the iteration process, we use only the top t keywords. If we use a tight threshold t in the initial steps of the iteration, the keywords and the retrieved documents will stay very specific to the original query. On the other hand, if we use a loose threshold, the keywords will become general ones within a few iteration. The threshold t can be used as a control parameter to specify the feature to be extracted.

We realized a search engine of the bankruptcy information based on the proposed method. We evaluated the method by comparing the extracted features with the correct words by human and confirmed that the method is effective to obtain specific feature depending on the purpose.

II. Related Work

A. Topic drift

There have been many text mining researches on extraction of important words and sentences. [20] clusters the sentences using the feature words, and choses the sentences which contains centroid as important sentence in the cluster. [5], [6] constructs the formal concept lattice from the documents of bankruptcy information and extracts the bankruptcy causes from the directed graph of the concept lattice. [13] and [18] constructs a network of words and proposed the TextRank method to evaluate the score of words as the HITS algorithm does for the authority-hub graph of Web pages. [17] proposed the bootstrapping method "Espresso" from the view point of text summarization and feature extraction.

These researches started independently. But, they are the same approach in the sense that they are analyzing the graph of word co-occurrence relation[19], [11]. In all of these researches, they consider the bipartite graph of words and documents or words and sentence and are extracting the dense subraph of the bipartite graph. Iterated traverse of the graph plays a key role in discovering feature words and important sentences. A large number of sentences that contain the same general words cause the long tail effect and the weight of such general words gain high score. The points of the research are in how to protect the drift. In this paper, we consider that a drift is avoidable by cutting off the propagation at each step of repetition.

B. Analysis of search result

A search engine is indispensable in order to discover required information from a huge amount of documents. But we often face a situation where a simple search is

not enough. In such a case, an exhaustive search and the overview of the whole result are required.

Survey of related work in academic research and consumer reports based on reputation information in blogs are typical samples of such task. Compiling a list of documents is not enough. Summary is required as well as characteristic words that describe the overview. There are many researches [1], [2], [14], [15], [24], [25] in text summarization and feature extraction which deal with this problem. It is known that characteristic words and phrase are effective to analyze when the target documents are restricted to the scientific articles, the reputation information or FAQs [4], [8]. [12] considers to discover such characteristic phrases in opposite way.

The research of text mining began in extraction of NE(named entities) such as name of person or name of place and has been developed further to text summarization and understanding. Deep analysis of the texts, such as the sentiment analysis of blogs [16], [7] and the causal analysis of incidents and financial market [9], [8], [21], [23], [10], [22] are attracting attention.

The method we propose in the present paper does not depend on such pattern or phrase. It is possible to integrate such methods with ours. We evaluate the effectiveness of our bootstrapping method to extract the bankruptcy cause.

III. Algorithm

Given a query, the proposed algorithm extracts the characteristic words of the search results. It is based on the index of documents and the index of sentences and constructed from the following four functions:

- $SearchDocuments(w)$ returns the set of documents which contain the query word w.
- $Sentence(d)$ returns the set of sentences in a document d.
- $SearchSentence(w)$ returns the set of sentences which contain the word w.
- $Score(u, S)$ returns the score of a word u in a set S of sentences.

The algorithm starts by applying the SearchDocuments, then iterates SearchSentences and TopKkeywords for fixed number of times as Fig.1.

IV. Evaluation with Ranking and Frequency

In this paper, we analyzed the bankruptcy information of Japanese companies publicly available at Tokyo Shoko Research [1]. Each document explains the profile of the bankrupt company and how the company went bankrupt. We constructed a search engine to analyze 431 documents ranging from March 2004 to July 2010. We separated the documents into 4024 sentences. We classified the sentences into the "cause sentences" and the non-cause sentences. We

[1]http://www.tsr-net.co.jp/

```
main(query){
  input w:keyword, t:threshold
  output FW:orderd list of characteristic words
  D = search-documents(w)
  # the documents that contain w
  S = sentence(D)
  # the set of sentences in the document D
  T = Bootstrap(S,t)
  # the set of sentences obtained by the algorithm;
  FW = TopKkeywords(t,T) # top t words in T
  return FW
}

Bootstrap(S,t){
  input S:set of documents, t:threshold
  output S:set of sentences
  for(i=0;i<max interation;i++){
    W = top-keywords(t,S)
    S = search-sentences(W)
  }
  return S
}
```

Figure 1. Restricted bootstrapping

focused the word "but" to determine if a sentence describes any cause of the bankruptcy or not and used the case sentences to extract the bankruptcy cause. Each documents, which contains 11.8 sentences in average, were analyzed manually and 13.1 words were chosen as the appropriate words for the bankruptcy cause of the company. We used these words as the correct answer to evaluate the result of the algorithm.

The proposed algorithm uses top t keywords for the document retrieval in the bootstrapping process. The effect of the threshold t is measured by comparing the result keywords with the correct cause words chosen manually. The process is iterated 10 times which is confirmed the convergence of the scores.

A. Feature words of "construction"

Table VI shows the ranked list of feature words of "construction" with respect to the threshold $t = 1, 2, 3, \cdots, 10, 20, \cdots, 90$. Top 10 words are shown in each line. The first column shows the threshold t, the second column shows the number of sentences, the third column shows the number of all words and the fourth column shows the rank of the query word "construction".

The words in the lines above $t = 4$ remind us "enforcement of the revised building standard law". So, we can say that the restriction on t worked effective in preventing the topic drift. On thee other hand, in the lines below $t = 50$, general words appear, such as reduction, collapse, downturn, deterioration and drop. At the same time, the ranking of the query "construction" is going down greatly. We confirmed that other queries yield the same ranked lists for $t = 90$. These analysis implies the effectiveness of the restriction on

t in preventing topic drift.

B. Ranking of word and frequency of word

From Table VI, we can say that general words are ranked at the top positions even if they are not tightly related to the query if the threshold t is loose, that is t is large. On the other hand, strongly related words are ranked at the top positions when the threshold t is tight, that is t is small.

Next we check if the top ranked feature words are not too general. Fig. 2 plots the frequency $s(w)$ of a word w in y-axis and the ranking $rank(w)$ of the word in x-axis. Note that we use the number of sentences that contain the word w as the frequency and that the frequencies are shown in log-scale. We can see that the lines with small t locate below compared to those with large t. This means that the top ranked keywords are rare and specific words which appear only in the related documents. On the other hand, the ploted lines with large t lie along the diagonal. This means that the words with high ranking are general words which occur in many sentences. From these analysis, the restriction of t can control the topic drift.

V. Evaluation as Bankuptcy Cause

In this section, we evaluate if the result words are correct as bankrupt cause with respect to the query.

As a evaluation measurement, we use MAP(Mean Average Precision). MAP is calculated as the average of average precision $AP(i)$. $AP(i)$ is the average of the precision for each outputs below rank i and is obtained as the number of correct guess divided by i. MAP is the average of $AP(i)$ when i varies from 1 to n. The value of MAP is better if the method gives correct guess in top rankings.

The correct answer of the bankrupt causes are manually chosen for each company. To evaluate the effect of restricting the threshold t, we selected the top 10 words whose frequencies are not larger than a parameter f. Thus, the $MAP(t, f)$ is calculated with respect to the two parameter t and f (Table V).

From this table, we can see that the smaller thresholds t yield better MAP. If we focus on the columns of $f \leq 10$ and $f \leq 20$, there is a clear gap between the smaller threshold t and the larger ones. This implies that the tight restriction on t extracts the specific feature with respect to the query. Table I and Table II show the top 10 words for $MAP(1, 10)$ and $MAP(50, 10)$. We can see that the keywords in Table I have lower frequency and are related to construction. On the other hand, the words in Table II have higher frequencies and are general words of bankruptcy which are not necessarily related to construction.

If we see the columns with larger f, the differene of MAPs according to t is small. Table III and Table IV display the top 10 words for $MAP(1, 500)$ and $MAP(50, 500)$. By comparing the two tables, we can see that the algorithm generates completely different keywords with $t = 1$ and $t = 50$, even though the MAP values are very close. The keywords with $t = 1$ have lower frequencies. The keywords with $t = 50$ have higher frequencies.

i	#ok	AP_i	weight		feature	DF
1	0	0.0000	8.4890	x	construction	14
2	0	0.0000	5.9255	x	standard	17
3	0	0.0000	5.6524	x	revision	19
4	0	0.0000	2.6110	x	enforcement	0
5	1	0.2000	2.5948	√	confirm	3
6	2	0.3333	2.3870	√	materials	5
7	2	0.2857	2.2309	x	rise	12
8	3	0.3750	2.1182	√	increase	5
9	4	0.4444	1.9020	√	long term	4
10	4	0.4000	1.7579	x	law	21

Table I
Top 10 features ($t = 1, f \leq 10, MAP = 0.2038$)

i	#ok	AP_i	weight		feature	DF
1	0	0.0000	24.5920	x	excess	26
2	0	0.0000	24.2462	x	deterioration	45
3	0	0.0000	22.8142	x	reduction	98
4	1	0.2500	22.6171	√	deficit	4
5	1	0.2000	22.3652	x	fall	0
6	1	0.1667	21.6223	x	fund	68
7	1	0.1429	21.5426	x	collapse	0
8	1	0.1250	21.4515	x	burden	74
9	1	0.1111	21.2447	x	downturn	35
10	2	0.2000	20.9756	√	large sum	8

Table II
Top 10 features ($t = 50, f \leq 10, MAP = 0.1196$)

We confirmed the similar situation for the other queries. Thus, we can say that the restriction on t is effective to extract the keywords tightly related to the query.

VI. Conclusion and further work

In this paper, we applied the bootstrapping method for feature extraction as a repetitive search process in the bipartite graph of sentences and words. We introduced a restriction on the frequency of the words to be not larger than t, in each step of iteration. A search engine was constructed for bankruptcy information to extract cause of bankruptcy

i	#ok	AP_i	weight		feature	DF
1	1	1.0000	8.4890	√	construction	14
2	2	1.0000	5.9255	√	standard	17
3	3	1.0000	5.6524	√	revision	19
4	3	0.7500	2.6110	x	enforcement	0
5	4	0.8000	2.5948	√	confirm	3
6	5	0.8333	2.3870	√	materials	5
7	6	0.8571	2.2309	√	rise	12
8	7	0.8750	2.1182	√	increase	5
9	8	0.8889	1.9020	√	long term	4
10	9	0.9000	1.7579	√	law	21

Table III
Top 10 features ($t = 1, f \leq 500, MAP = 0.8904$)

i	#ok	AP_i	weight		feature	DF
1	1	1.0000	24.5920	$\surd$	excess	26
2	2	1.0000	24.2462	$\surd$	deterioration	45
3	3	1.0000	22.8142	$\surd$	reduction	98
4	4	1.0000	22.6171	$\surd$	deficit	4
5	4	0.8000	22.3652	x	fall	0
6	5	0.8333	21.6223	$\surd$	fund	68
7	5	0.7143	21.5426	x	collapse	0
8	6	0.7500	21.4515	$\surd$	burden	74
9	7	0.7778	21.2447	$\surd$	downturn	35
10	8	0.8000	20.9756	$\surd$	large sum	8

Table IV
TOP 10 FEATURES ($t = 50, f \leq 500, MAP = 0.8675$)

for an arbitrary condition. Effectiveness of the proposed method is confirmed by evaluating the MAP of ranked list of feature words obtained. It was confirmed that the infrequent words that are tightly related to the query can be extracted by restricting the threshold t to be small.

We use the word "but" as clue to specify the cause-sentences. There might be other candidates as clue words. The search and the comparison of clue words is one of the further work. It will be interesting and possible to find similar clue words for the scientific literatur by which the aim and the purpose of the paper should be extracted. The evaluation was conducted using GETA[2] search engine and we used "SMART" weight as the score of the keyword. It is worthwhile to check the result with other measure.

REFERENCES

[1] R. Brandow, K. Mitzeb, L. F. Rau, Automatic condensation of electronic publications by sentence selection, Information Processing & Management Vol. 31, No. 5, pp. 675–685, 1995

[2] R. Bhowmik, Keyword Extraction from Abstracts and Titles, Proc. IEEE Southeastcon, pp.610–617, 2008

[3] R. Feldman, J. Sanger, Handbook of Natural Language Processing, Cambridge University Press, 2007

[4] R. Higashinaka, H. Isozaki, Automatically Acquiring Causal Expression Patterns from Relation-annotated Corpora to Improve Question Answering for why-Questions, ACM Trans. on Asian Language Information Processing, Special Issue on NTCIR-6, Volume 7, Issue 2, Article 6 (29 pages), 2008.

[5] S. Hirokawa, T. Baba, T. Nakatoh, Text Mining of Bankruptcy Information using Formal Concept Analysis, Proc. 3rd. ICAST, 2011 (to appear)

[6] S. Hirokawa, T. Baba, T. Nakatoh, Search and Analysis of Bankruptcy Cause by Classification Network, Springer LNCS, Vol 6918, pp.152–161, 2011

[7] D. Ikeda, H. Takamura, M. Okumura, Learning to Shift the Polarity of Words for Sentiment Classification, Journal of JSAI, Vol.25, No.1, pp.50–57, 2010 (in Japanese)

[2]http://geta.ex.nii.ac.jp/geta.html

[8] T. Inui, M. Okumura, Characteristics of In-text Causal Relations, Tech. Report IPSJ 2005-LN-167, 2005 (in Japanese)

[9] H. Isozaki, R. Higashinaka, Anwering System for "why" using Patter Mining, Proc. Ann. Conf. JNLP, pp.1025-1028,2008 (in Japanese)

[10] K. Izumi, T. Goto, T. Matsui, Analysis of Financial Markets Fluctuation by Textual Information Journal of JSAI, Vol. 25, No.3, pp.383-387, 2010 (in Japanese)

[11] M. Komachi, T. Kudo, M. Shimbo, Y. Matsumoto, Semantic Drift in Espresso-style Bootstrapping: Graph-theoretic Analysis and Evaluation in Word Sense Disambiguation, Journal of JSAI, Vol.25, No.2, pp. 233–242, 2010 (in Japanese)

[12] M. Krapivin, A. Autayeu, M. Marchese, E. Blanzieri, N. Segata, Keyphrases Extraction from Scientific Documents: Improving Machine Learning Approaches with Natural Language Processing Springer LNCS 6102, pp.102-111, 2010

[13] R. Mihalcea, P. Tarau, TextRank:Bringing Order into Texts, In Proceedings of EMNLP'2004. pp.404–411, 2004

[14] M. Murata. S. De Saeger, C. Hashimoto, J. Kazama, I. Yamada, K. Kuroda, Extracting and visualizing important information from article abstracts, Proc. Ann. Conf. JSAI, 2009

[15] T. Nagai, K. Kaneko, H. Shibuki, M. Nakano, R. Miyazaki, M. Ishige, T. Mori, Important Passage Extraction based on Muli-Layered TextRank Network, Proc. Ann. Conf. ANLP, pp. 294–297, 2010 (in Japanese)

[16] aT. Nasukara, H. Kanayama, Acquisition of Sentiment Lexicon by Using Context Coherence, Tech. Report IPSJ, 2004-73, pp. 109-116, 2004 (in Japanese)

[17] P. Pantel, M. Pennacchiotti, Espresso: Leveraging Generic Patterns for Automatically Harvesting Semantic Relations, Proc. 44th ACL, pp.113-120, 2006

[18] G. K. Palshikar, Keyword Extraction from a Single Document Using Centrality Measures, Springer LNCS 4815, pp. 503–510, 2007

[19] D. Radev, R. Mihalcea, Network and Natural Language Processing, AI Magazine, pp.16-28, 2008

[20] D. R. Radev, H. Jing, M. Stys, D. Tam, Centroid-based summarization of multiple documents, Information Processing and Management, Vol. 40, pp.919–938, 2004

[21] U. Shibusawa, T. Hayashi, R. Onai, RE:Why-a Prototype System for Extracting Answers of a "Why" type Question from Web Pages, Computer Software, Vol.24, No.3, pp.20-28, 2007 (in Japanese)

[22] H. Takeuchi, S. Hagino, H. Watanabe, K. Shirota, Bankruptcy Analysis by Text Mining, Proc. Ann. Conf. JSMI, 2008 (in Japanese)

[23] K. Tanaka, T. Takimoto, Y. Kiyoki, Domain Independent Why Text Segment Classification and Answer Extraction by Grammer-gram and Grammar Verg-gram, Tech. Rep. IEICE WI2, pp.89–94, 2009 (in Japanese)

	f									
t	$\leq$10	$\leq$20	$\leq$30	$\leq$40	$\leq$50	$\leq$100	$\leq$200	$\leq$300	4$\leq$00	$\leq$500
1	0.2038	0.8804	0.8904	0.8904	0.8904	0.8904	0.8904	0.8904	0.8904	0.8904
2	0.1591	0.8357	0.8693	0.8693	0.8693	0.8693	0.8693	0.8693	0.8693	0.8693
3	0.1325	0.8707	0.9043	0.9043	0.9043	0.9043	0.9043	0.9043	0.9043	0.9043
4	0.1325	0.8257	0.8593	0.8593	0.8593	0.8593	0.8593	0.8593	0.8593	0.8593
5	0.1425	0.8807	0.9453	0.9453	0.9453	0.9453	0.9453	0.9453	0.9453	0.9453
10	0.1929	0.2508	0.2719	0.2719	0.5648	0.5648	0.5648	0.5648	0.5648	0.5648
20	0.1929	0.3254	0.3254	0.3254	0.6394	0.7823	0.7823	0.7823	0.7823	0.7823
30	0.0846	0.0846	0.3775	0.4420	0.6349	0.8468	0.8468	0.8468	0.8468	0.8468
40	0.1196	0.1196	0.4125	0.4461	0.6390	0.8509	0.8509	0.8509	0.8509	0.8509
50	0.1196	0.1196	0.4125	0.4336	0.6265	0.8675	0.8675	0.8675	0.8675	0.8675

Table V
MAP WITH RESPECT TO (t, f)

Table VI
RANKED LIST OF FEATURE WORDS OF "CONSTRUCTION"

t	S	W	R	top 10 words
1	22	341	1	construction standard revision enforcement confirm rise materials "long term" law influence
2	22	379	3	revision standard construction "industrial law" rise law influence enforcement confirm materials
3	32	482	3	revision standard construction rise enforcement confirm materials law influence groundbreaking
4	88	812	4	"market condition" revision standard construction "subprime loan" deterioration rapid beginning problem downturn
5	111	952	19	"market condition" problem "market condition" beginning rapid emit revision deterioration emit environment
6	193	1257	88	"market condition" deterioration problem rapid beginning environment emit emit market
7	206	1323	90	"market condition" deterioration problem rapid beginning environment emit emit market
8	206	1323	90	"market condition" deterioration problem rapid beginning environment emit emit market
9	244	1462	62	"market condition"deterioration problem rapid beginning environment influence emit
10	249	1483	35	"market condition" deterioration problem rapid environment beginning influence revision emit
20	464	2077	125	collapse deterioration "market condition" reduction downturn bubble drop "market condition" problem rapid
30	541	2321	153	reduction collapse downturn deterioration drop "market condition" achievements deficit bubble fall
40	609	2500	170	reduction collapse downturn deterioration drop "market condition" burden deficit achievements bubble
50	661	2686	197	reduction collapse downturn deterioration drop "market condition" excess burden deficit achievements
60	714	2805	201	reduction collapse downturn deterioration drop burden excess "market condition" deficit bubble
70	754	2899	198	reduction collapse deterioration downturn drop burden "market condition" excess "large sum" deficit
80	787	2973	218	reduction collapse deterioration downturn drop burden "market condition" excess "large sum" deficit
90	798	3012	204	reduction collapse deterioration downturn drop burden "market condition" "large sum" excess deficit

[24] S. Teufel, M. Moens, Summarizing Scientific Articles - Experiments with Relevance and Rhetorical Status Computational Liguistics, Vol.28, Iss. 4, pp. 409–445, 2002

[25] P. Turney, Learning Algorithms for Keyphrase Extraction, Information Retrieval Journal, 2(4), pp.303–336, 2000

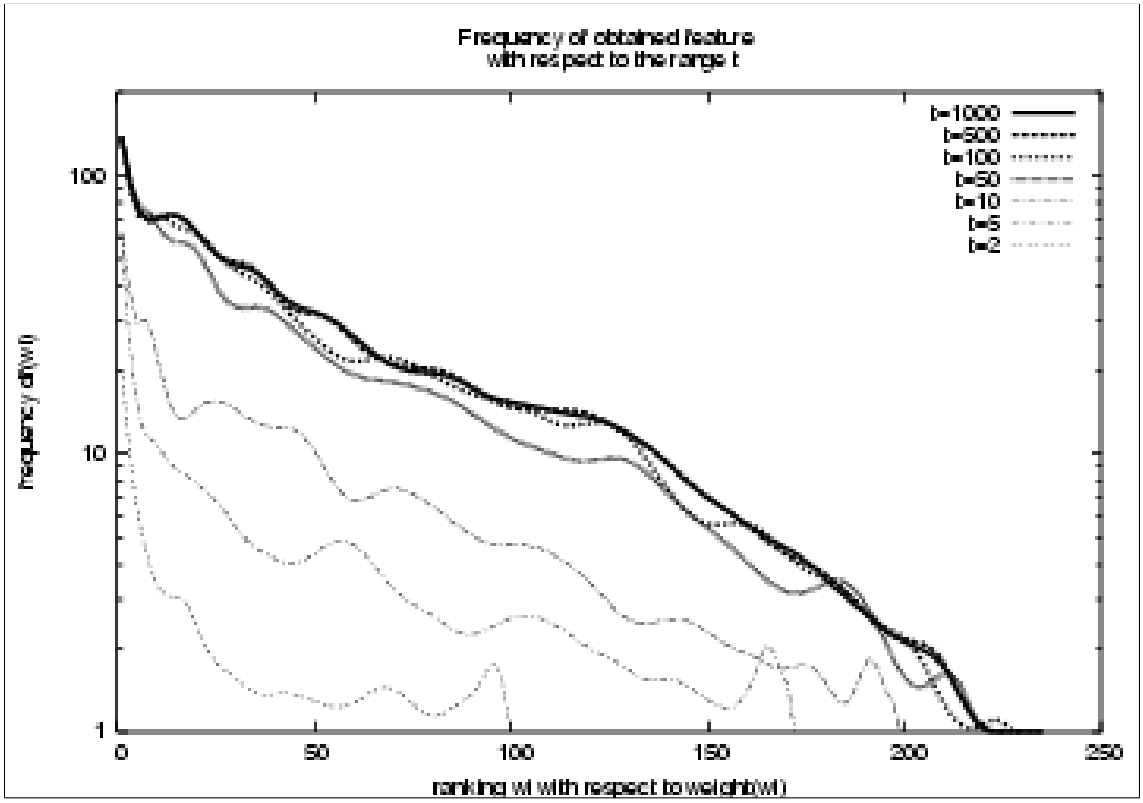

Figure 2. Document frequencies of keywords