Statistical Approach to Estimate the Quality of Web Datasets

Vitaly KLYUEV
Software Engineering Laboratory, University of Aizu
Aizu-Wakamatsu City, Fukushima 965-8580, Japan

ABSTRACT

Finding appropriate information on the Web is getting more difficult with inefficient tools currently being used on the net. Using a topic-specific approach to build crawlers is promising. In this paper, we discuss a technique using methods of statistical analysis to evaluate the quality of the crawled documents. We have found this technique is more robust, more reliable, more practical and less subjective compared to others.

Keywords: Search Engine, Similarity Metrics, Focused Crawler.

1. ITRODUCTION

The role of search engines as tools to find appropriate information increases, because without being indexed, data cannot be found and accessible. On the other hand, the more powerful search engines are, the smaller percentage of data from the net can be indexed by them. The rate of improving search tools is much less than the rate of the data growth on the net. We agree with a note made in [6]: Searching is not only one of the most common tasks performed on the Web but one of the most frustrating.

Maybe it is time to stop thinking about a catalogue of documents like the Yahoo catalogue on the net. We need to take a step forward: the time for catalogues of topic-specific servers has come. General purpose search engines like Google and AlltheWeb are playing a significant role in finding starting points on the net to search for detailed information to answer user queries. To make these steps easier for people, topic-specific search engines can be helpful. These search engines have to be tools to sort out the chaotic world of electronic documents on the net.

A semi-automatic style to create collections is natural. The current level of information technologies is good enough to provide researchers with the necessary methods and tools. The key issue is how to evaluate the quality of crawled collections.

In this paper, we discuss one of the possible ways to do that. Our approach is to use the methods of statistical analysis.

The rest of the paper is organized as follows. The next section gives a short overview of related work. Section 3 is devoted to evaluation of the quality of the collections' content. It presents statistical analysis of obtained results. Final remarks and our plans for future work are in the Conclusion section.

2. RELATED WORK

Focused crawlers usually filter an input stream to select documents relevant to the topic of interest. Several approaches were proposed to design a focused crawler; some of them can be seen in [2, 3, 7, 8, 9, and 13]. All of the approaches use heuristics based on the intuition what related documents connected by links and links extracted from one document will follow to semantically close ones. Methods differ from each other in the following: What the filter is and how to detect a relevant document. Most of the techniques adopt the idea of Page Rank algorithm [1].

What is the problem with "understanding" documents by computers? Many approaches [7] take into account a vision of any document as a bag of words and compounds. If computers index a program code, we lose everything doing this kind of indexing. Examples from Table 1 illustrate the main problem of information retrieval, which is in the lack of a language model. As it was noted in [6], context extraction is far less practical, and because of this, none of the search techniques is able to deal effectively and efficiently with the huge volume of information growing on the net.

Nowadays, a natural language processing technique

Table 1 Document Indexing

Source (text)	Index	Source (program code)	Index

According to several estimations, the amount of data created in the last two years is as big as the accumulated data in all human history. This exponential growth trend continues.	according 1
	several 1
	estimation 1
	amount 1
	data 2
	create 1
	…
	continue 1
		static boolean isPrime(int n){
		if (n <= 2)
		{ return n == 2; }
		if (n % 2 == 0)
		{ return false; }
		for (int i = 3, end = (int)Math.sqrt(n);
		i <= end; i += 2) {
		if (n % i == 0) { return false; }
		}
		return true;
		}
			static 1
			boolean 1
			isPrime 1
			int 3
			n 5
			{ 5
			if 3
			<= 2
			2 4
			…
			} 5

can provide researchers only with morphological analyzers. They are helpful in segmenting texts in Asian languages and classifying components of sentences in many languages.

How do authors of alternative approaches evaluate their systems?

In study [11], they use 22 questions to be submitted to the designed system and to two commercial ones. Authors compare the retrieval results. There are at least two points raise doubts: a) the systems retrieve documents from completely different databases, b) as we pointed out, mechanisms to crawl data and to retrieve documents are different. An evaluation of crawled mechanisms on the base of the quality of retrieved data is highly suspect.

Authors of the approach presented in [10] do not create any collection in advance. They designed a front-end system for a commercial search engine and tested it in the Japanese gastronomy domain. The evaluation on the base of the quality of retrieval is justified.

Study [3] provides a discussion about algorithms to crawl the net only. It did not estimate the quality of collections created.

Authors of [2] evaluated their approach running the designed crawler several times with different parameters and using a randomly selected part of the initial seed of URLs. They estimated the percentage of common pages crawled, etc. There is no relation between the percentage of common pages and the number of relevant pages found.

As we can see, an evaluation of the crawled content is a very difficult task. We agree with authors [2] that it is extremely difficult to measure or even define parameters such as recall for a focused crawler, because we have a rather incomplete and subjective notion of what is good coverage on a topic.

Specifics of compiling document collections from the Internet should be taken into account when one evaluates the quality of obtained data. They are as follows:

1) Web crawl cannot be reproduced in the same way even by the same system. This is because the Internet is constantly changing. It means, the same system will not be able to compile the same content even in a short period of time.
2) As we mentioned earlier, approaches use heuristics based on human intuition. There is not a good language model to be applied to retrieval systems.
3) There are a myriad of techniques for compiling collections as well as retrieving documents from them prior to presenting the documents to the end user. On the other hand, we can see and estimate the result of retrieval as a response of the system to the user query.
4) In many cases, it is difficult to distinguish between relevant and non relevant documents to the topic of interest.

What can we learn from the experience of the famous scientific forums like TREC, CLEF, and NTCIR? They recently started running a Web retrieval task. To make a comparison possible, different approaches have to be applied to the same data set. Participants are provided with a huge data set (100 Gb to 1Tb) crawled from the Web. They are requested to retrieve relevant documents to the given queries as well. This task differs from ours.

To evaluate retrieval results, they use a pooling method [4]. According to this method, they merge high-ranked results submitted by participants in one pool. Human assessors judge the relevance of each document in this pool. They classify documents as:

a) highly relevant,
b) fairly relevant,
c) partially relevant and
d) irrelevant.

According to study [12], the number of assessors has to be in the range of 2 to 4. We note here that the understanding of relevance is highly subjective. If several assessors have different opinions about the same document and query, it becomes unclear how to set up the final judgment.

3. EVALUATION OF THE COLLECTIONS

We selected English and Japanese scientific collections [5, 6] to evaluate the quality of the crawled documents using the methods of statistical analysis [12]. The reason why we used them is: They were analyzed by human inspection in detail. The size of our collections is as follows: The Japanese collection includes 64,074 documents; the English collection consists of 16,242 documents.

The sign test was selected among nonparametric tests because it does not require any assumption about the shape of the population involved. The second evaluation was done making the following assumption: the conditions for a binomial experiment were satisfied. The claim about proportion was tested.

The same sample data set was used in both tests. It was obtained as follows: 1000 documents were randomly selected from each set: Japanese documents (64,074) and English pages (16,242). Every document was manually estimated and classified as relevant or non relevant. Results are presented in Table 4. See it on the next page. The number of broken links illustrates the changeable nature of the Internet: 6.6% of them could not be reached. The analysis of Japanese set was conducted one month after finishing the crawl.

We should note the estimation of the crawled data and its quality. What is a good document? Is there a border between relevant and non-relevant documents? These questions are important in the topic specific search engine area.

Table 4 Characteristics of randomly selected data

Parameter / Collection	Japanese	English
Number of relevant documents	503	644
Number of non relevant documents	431	356
Number of broken links	66	
Number of hosts	283	347

The main aim was to create a large collection of algorithms and their applications which could be used in teaching in schools and universities. Our point of view is as follows: 1) documents including a description of any algorithm were marked as relevant; 2) documents which could be used in finding documents with algorithm descriptions were also marked as relevant (for example, home page of any university computer department). Even making this assumption, it was a subjective decision about the relevance of documents in many cases. This note should be taken into account when assessing data in Table 4.

3.1. Sign Test

Japanese data All relevant documents from Table 4 were denoted by plus and non-relevant ones by minus. The number of samples is N=934. We subtracted 66 from the total number of samples. Our claim is that 50% of documents among the crawled set are non-relevant to our topic of interest.

Null and alternative hypotheses are

- H0: P=0.5, the number of non-relevant documents is half of the crawled amount,
- H1: P<>0.5

Is the conflict with the null hypothesis significant? N= 934 and the number of non-relevant documents in our sample is equal to 431. Because N=934 and its value is more then 25, the test statistic z is based on a normal approximation to the binomial probability distribution with p=q=0.5. The test statistic z is as follows:

$$z = \frac{(x+0.5)-(N/2)}{\sqrt{N}/2} = -2.32$$

Here x equals the number of the less frequent sign (x=431). With a significance level of 0.05 in a two-tailed test, the critical values are z=+/-1.96. The test

statistic is less than these critical values, so we should reject the null hypothesis of equality.
From this test, we can conclude that the number of relevant documents in our collection is more than half.

English data Following the aforementioned procedure and applying data from Table 4, we should reject the null hypothesis of equality as well.

3.2. Claim About Proportion

Japanese data The following assumptions are fulfilled:

1) The conditions for a binomial experiment are satisfied. We have a fixed number of independent trials, which have constant probabilities. Each trial has two outcome categories: relevant and non relevant.
2) The conditions N*p>= 5 and N*q>=5 are also satisfied (discussed later).

Following this, we can apply methods of parametric statistics.

Null and alternative hypotheses are

- H0: P>=0.56, the number of relevant documents is at least 56% of the total amount of the crawled data,
- H1: P<0.56

In this case, the value of the test statistic z is as follows:

$$z = \frac{p^1 - p}{\sqrt{\frac{p * q}{N}}} = -1.25$$

Where N is a number of trials (934 in our case), p is a population proportion (given in the null hypothesis), p^1 = x/N = 503 / 934 = 0.54 (sample proportion), and q=1 – p.
With a significance level of 0.05 in a one-tailed test, the critical value of z= -1.645. Because the test statistic does not fall within the critical region, we fail to reject the null hypothesis.

The main result from this is as follows: The proportion of relevant documents in our collection is at least 56%.

English data Null and alternative hypotheses are

- H0: P>=0.66, the number of relevant documents is at least 64% of the total amount of the crawled data,
- H1: P<0.66

In this case, the value of the test statistic is z= -1.32 and we fail to reject the null hypothesis. The proportion of relevant documents in this collection is 10% higher compared to Japanese one. This result is in accord with our manual analysis [7].

4. CONCLUSION AND FUTURE WORK

The main point of our proposal in this paper is to use statistical methods to evaluate the quality of crawled document collections compiled using the different techniques. Results of our statistical analysis correspond to results of manual analysis of data sets tested. This approach can be applied to estimate the full set of documents before indexing and incorporating into an index of a search engine. This method is more robust, more reliable, more practical and less subjective compared to others.
The next task in our research is to minimize the subjectivity at the step of manual analysis of sample data.

4. ACKNOWLEDGEMENT

This research was funded by Fukushima Prefectural Foundation for Advancement of Science and Education, grant F-11, “Compiling and keeping up-to-date large topic-specific data from the Internet”, 2003.

5. REFERENCES

[1] Sergey Brin and Lawrence Page, “The Anatomy of a Large-Scale Hypertextual Web Search Engine”, **Computer Networks and ISDN Systems**, 29(11):1257-1267, 1997.

[2] Soumen Chakrabarti, Martin van den Berg and Byron Dom, “Focused Crawling: A New Approach to Topic-Specific Web Resource Discovery”, **Proc. of 8th International World Wide Web Conference (WWW8)**, 1999.

[3] Michael Chau and Hsinchun Chen, “Personalized and Focused Web Spiders”, **Web Intelligence**, Springer Verlag, 2003, pp. 197 – 213.

[4] Koji Eguchi, Keizo Oyama, Akiko Aizawa and Haruko Ishikawa, “Overview of the Information Retrieval Talk at NTCIR-4

WEB", **Proc. of the NTCIR Workshop 4 Meeting. Working Notes of the Fourth NTCIR Workshop Meeting**, (Supplement Volume 1, 2), NII, Tokyo, Japan, June 2 -4, 2004, pp. ov3 – ov15.

[5] V. Kluev, "The Core of a Topic-Specific Search Engine: How to Create It", **The WSEAS Transactions on Communications**, Issue 1, Vol.3, January 2004, pp. 188—192.

[6] Wen-Chen Hu and Jyh-Haw Yeh, "World Wide Web Search Engines", **Architectural Issues of Web-Enabled Electronic Business**, Idea Group Pub, 2002, pp. 154 – 169.

[7] V. Kluev, "Compiling Document Collection from the Internet", **ACM SIGIR Forum**, Vol. 34, Number 2, pp. 9 – 14, Fall 2000.

[8] V. Kluev, "Results Merging with the OASIS System: An Experimental Comparison of Two Techniques", **IEICE Transactions on Information and systems**, VOL.ED86-D,No. 9, September 2003, pp.1773 – 1780.

[9] A. McCallum, K. Nigam, J. Rennie, and K. Seymore, "A Machine Learning Approach to Building Domain-Specific Search Engines", **Proc. Of the Sixteenth International Joint Conference on Artificial Intelligence (IJCAI-99**), 1999.

[10] Satoshi Oyama, Takashi Kokubo, Teruhiro Yamada, Yasuhiko Kitamura and Toru Ishida. "Keyword Spices: A New Method for Building Domain-Specific Web Search Engines", **International Joint Conference on Artificial Intelligence (IJCAI-01)**, pp.1457-1463, 2001.

[11] Jialun Qin, Yilu Zhou, and Michael Chau, "Building domain-specific web collections for scientific digital libraries: a meta-search enhanced focused crawling method", **Proc. of the 2004 joint ACM/IEEE conference on Digital libraries**, Tucson, AZ, June 7 - 11, 2004.

[12] Mario F. Triola, **Elementary Statistics**, Addison-Wesley Publishing Company, 1995, 726 p.

[13] Ah Chung Tsoi, Daniele Forsali, Marco Gori, Markus Hagenbuchner, and Franco Scarselli, "A Simple Focused Crawler", **Proc. Of the Twelfth World Wide Web Conference**, *Hungary*, 2003.