COPYRIGHT © 2010 JCIT, ISSN 2078-5828 (PRINT), ISSN 2218-5224 (ONLINE), VOLUME 01, ISSUE 01, MANUSCRIPT CODE: 100708

A Novel Bangla Text Corpus Building Method for Efficient Information Retrieval

Jubayer Shamshed and S. M. Masud Karim

Abstract—This paper presents a novel and effective method for building Bangla text corpus in order to allow efficient information retrieval. Information retrieval system is yet not used for Bangla text. Nowadays Random walk algorhithm have got good score for information retrieval system research. And priority base term frequency gives extra weight in different part of the text, like title, abstract, author, body of text etc. In our paper, terms/words in the title have been given extra frequency; because the terms/words in title are considered very important for information retrieval system. The precission and recall measure the effectiveness of the retrieved system. The underlying formulation and the obtained results show that the proposed method result in Bangla text corpus and the performance of information retrieval is as good as any existing text corpus in terms of precission, recall and zipf's curve.

Index Terms— Bangla text corpus, Information retrieval, Metadata, Random walk algorithm.

1 Introduction

Information retrieval (IR) means retrieving information from the database or collection of documents. Information Retrieval and Summarization is one of the most successful steps in Natural Language Processing (NLP) and one of the key steps for information extraction technology particularly in the domain of World Wide Web (WWW) information storage, query processing, and answer extractions etc. The automated IR systems have already started providing efficiency in contrast with human-based IR systems as earlier mentioned systems are proved to be much quicker and less expensive. The IR is an extremely broad field, cncompassing a wide range of topics pertaining to the storage, analysis and retrieval of all manner of media. But these IR systems are still improvising a lot of research interests to researchers as natural language contains inherent ambiguities and much of the challenge of NLP invovles resolving these ambiguities. Corpus building and corpus-based IR systems are drawing focus due to such challenges. A corpus is defined as a collection of large amount of texts. It is as important resource as any other resources for linguistic research. Brown corpus is one of the English corpuses for linguistic research [7]. Dutch corpus is well known to retrieve Dutch text [6]. A Persian corpus is built in [1] for Persian language retrieval system. But an effective Bangla corpus is hardly found for Bangla IR purposes.

There is a number of researches available describing the feature of the Bangla linguistic corpus. A well known Bangla corpus based on Indian Bangla language is presented in [11] and that corpus is effective for linguistic activities. A Bangla corpus based on newspaper is analyzed in [8]. A number of works are devoted to have the corpora on Bangla Optical Character Recognition (OCR). The contribution of [4] is also in the specific area of Bangla OCR. Besides, in [9], a method for building an effective corpus for evaluation of Bangla text compression is presented and that helps to build corpus. But this method is for text compression only. In this paper, a method for building Bangla text corpus is proposed which is effective for IR system. In addition, the criteria for evaluation of a corpus with respect to IR are also described.

- *Jubayer Shamshed is with the Computer Science and Engineering Discipline, Khulna University, Khulna 9208, Bangladesh. E-mail: j3shamshed@yahoo.com.*
- *S. M. Masud Karim is with the Computer Science and Engineering Discipline, Khulna University, Khulna 9208, Bangladesh. E-mail: masud@cse.ku.ac.bd.*

2 Literature Reviews

2.1 Corpus

In linguistics, a corpus (plural corpora) or text corpus is a large and structured set of texts, usually electronically stored and processed [12]. They are used to do statistical analysis and hypothesis testing, checking occurrences or validating linguistic rules on a specific universe. Query based IR can become effective if the information storage system gather vast amount of information. As the corpus has large amount of information, it can be used for effective retrieval. The characteristics of a corpus are defined by the purposes of creating the corpus and the evaluation criteria for which the corpus is designed for. These criteria give rise to the necessity to field specific corpora like IR, data compression, data mining etc. There are some authorized characteristics for corpus for IR. IR based corpus means that the corpus has the ability to retrieve effective desired information. There are a vast number of text documents in a text corpus.

2.2 Criteria for Corpus in IR

According to [8], there are six criteria for choosing a corpus. In IR process, basic criteria are: document specification, topics specification, query analysis, judgement on relevant or irrelevant document, evaluation, ranking the documents [1].

Document specification means dividing information into some categories such as title, abstract, body, bibliography etc.

Topics specification means description of the document (single sentence), title (important term in the corresponding document) and narrative.

Query analysis is based on manual and automatic operation in the preliminary stage of corpus for all documents. This set of query is called smart query.

Judgement on relevant or irrelevant documents is the most important consideration because in the mean time, IR system will be frustrated if the desired documents are not availalble or irrelevant. When valuable information is being looked for, many irrelevant documents are also retrieved and many relevant documents might be missing.

Paper [1] can be considered as a basis for corpus building for Asian languages. In this paper, the criteria like document specification, topics specification, and query analysis, judgement on relevant or irrelevant document, evaluation and ranking of documents are employed. In [6], incorporates a simplistic scheme for evaluating precision and recall, which are the key concerns for IR system. Paper [9] emphasizes on exploring a relationship between Type to Token Ratio (TTR) and compression ratio with maximum support for the performance evaluation of text compression. This paper considers ten groups namely, article, poems, advertisements, speeches, news, SMS, E-mails, stories and reports. The TTR is calculated for each file. BHASA is one of the search engines in Bangladesh to retrieve Bangla information based on Bangla corpus [10]. This search is perfomed on daily Bangla newspaper 'Prothom-Alo'. This Prothom Alo news corpus is analyzed in [8] using word level frequency analysis, TTR, average word length and zip's curve.

Zipf's Law tells us how much text we have to look at and how precise our statistics have to be to achieve what level of expected error. For example, the most frequent 150 words typically account for around half the words of a corpus. And Zipf's Law provides a base-line model for expected occurence of target terms and the answers to certain questions may provide considerable information about its role in the corpus. The law discovered empirically by Zipf (1949) for word tokens in a corpus states that, the frequency of word tokens in a large corpus of natural language is inversely proportional to the rank. That is, if f is the frequency of a word in a corpus and r is the rank, then

$$f = k / r \quad (1)$$

Where, k is a constant for the corpus [8]. When log(f) is drawn against log(r) in a graph, a straight line is obtained with a slope of –1. This line is called a Zipf curve.

Term frequency means how many times a terms frequently used in the text corpus [8]. But in IR system, title, abstract publication date, author names etc are more important than any other part of the documents. So, we consider priority based term frequency equation for calculating terms frequency. And remove stop-words. The equation is:

$$tf_{ijk} = \alpha \cdot tf_{ij1} + \alpha \cdot tf_{ij2} + \beta \cdot tf_{ij3} + \gamma \cdot tf_{ij4} \quad (2)$$

In [3], it is discussed the Random walk graph based ranking algorithms and its text rank adaption to drive term weights from textual graphs. Textual graphs encode term dependencies in text. This algorithm shows the highly important word in the text documents based on, which documents are very much connected with each one another. The base equation is:

$$s(v_i) = (1 - d) + d * \sum_{j \in in_{vi}} \frac{s(v_j)}{\left| Out \ (s(v_j)) \right|} \quad (3)$$

Random walk graph are considering both undirected and directed graph graph. Here we are using directed graph graph for random walk algorithm. In the context of Web surfing or citation analysis, it is unusual for a vertex to include multiple or partial links to another vertex, and hence the original definition for graph-based ranking algorithms is assuming un-weighted graphs. However, in our proposed model the graphs are build may include multiple or partial links between the units (vertices) that are extracted from text. It may be therefore useful to indicate and incorporate into the model the 'strength' of the connection between two vertices as a weight added to the corresponding edge that connects the two vertices.

Precision and recall are shown the effectiveness of IR system and efficiency of the corpus [6]. In [5], attention is focused on the effective standard deviation of precision and recall. Because recall measures how well a system retrieves all of the relevant documents and precission measures how well the system retrieves only the relevant documents.

The precission and recall are measuring by following equation:

$$precision = \frac{Number\ of\ relevant\ documents\ retrieved}{Total\ number\ of\ documents\ retrieved} \quad (4)$$

$$recall = \frac{Number\ of\ relevant\ documents\ retrieved}{Total\ number\ of\ relevant\ documents} \quad (5)$$

However, the problem of Prothom Alo news corpus is that it handles the storage of the newspaper only, but IR is not considered.

3 Proposed Corpus Building Method

This paper presents a method for building a successful Bangla text corpus that can be used for efficient IR system. The following steps are considered in building process of the corpus.

3.1 To Collect Documents

Documents are collected through different sources such as newspapers, books, WWW etc. The Documents are stored in the source folder as text files. The texts of the documents are converted into texts in Unicode. The Unicode fonts are

available on the net. The conversion is done using the 'Avro' Unicode converting software.

3.2 To Format Documents

Respective documents are collected from the source folder and a unique ID has been assigned to each document. These IDs are used as superkeys in the corpus. After that, the respective documents are formated into a division using five criteria, as follow:

1. Title (.T): Identifying the unique title of the document.
2. Categories (.C): Identifying the type of the contents in the document like poem, book, blog, political critisism etc.
3. Body (.B): The body of text has five different layers, such as:
    a. Abstract.
    b. Introduction.
    c. Main text
    d. Conclusion
    e. Bibliography
4. Author (.A): Mentioning the name(s) of the author(s) of the text.
5. Publication (.P): Marking the publication place and date of the text.

Example 3.1: Document formation

.ID 102
.T যাওয়া
.A আনিসুল হক
.C বই
.P ডিসেম্বর ০৬, ২০০৮
.B বাবা বললেন, শাহিন, অন্যরা যাচ্ছে যাক, তুই কিন্তু যাবি না□ না বাবা, যাব না□ অন্যের ছেলেরাও ছেলে□ অন্য বাবারাও বাবা□ তবু তোর যাওয়ার দরকার নাই? তুই কী বলিস!

3.3 To Calculate Distinct Words or Terms

A word or term may be used several times in a document, but in this calculation it will be counted once. Distinct word calculation discards the simultaneous use of a word and considers it as a single word.

Example 3.2: Suppose the body of a document is *a b c d e f a b b f z x*. Then distinct words are *a b c d e f z x* and number of distinct words in the document is 8.

3.4 To Claculate Term Frequency

After indexed the distinct words, the term frequency of each term in the text corpus is calculated. Then the term frequency is arranged in ascending order in a separate file with the records of ocurrences. The term frequency in different priority is calculated using the following equation:

$$tf_{ijk} = \alpha \cdot tf_{ij1} + \alpha \cdot tf_{ij2} + \beta \cdot tf_{ij3} + \gamma \cdot tf_{ij4} \quad (6)$$

Where,
α = 4, for title, author name.
β = 2, for categories.
γ = 1, for body of texts.
i = Document number.
j = number of term.
k = 1 for title, 2 for author name, 3 for categories and 4 for body of text

Example 3.3: Given the following document

.ID 19.
.T 'a' 'b' 'c'.
.A 'd'.
.B 'a' 'b' 'c' 'g' 'h' 'j' 'm'.

Term frequency for 'a' can be calculated as
$tf_{111} = 4 \cdot 1 + 4 \cdot 0 + 0 \cdot 0 + 1 \cdot 1 = 5$.
Calculation of term frequency is same for 'b', 'c', 'd' , 'g', 'h', 'j', and 'm'.

3.5 To Apply Random Walk Term Algorithm

The random walk computes the scoring of that term by using following algorithms and equation. So the weighting of terms calculated using above models does not reflect the actual weight of term. We solve the mentioned problem using the following steps:

- Remove stop-words from text documents [13].
- Build a graph
  # buildGraph
  Select aWord
  Add previousWord in inOrderd
  Add nextWord in outOrdered
  Assign initial value 0.25
  # endGraph
- The Equation is,

$$s(v_i) = (1-d) + d * \sum_{j \in in_{vi}} \frac{s(v_j)}{\left|Out\ (s(v_j))\right|} \quad (7)$$

Where,

$s(v_i)$ = Score or weight of vertex v_i.

$s(v_j)$ = Score or weight of vertex v_j.

$Out(s(v_j))$ = Out degree of vertex v_j.

d = Damping factor.

3.6 To Assemble Metadata

From formatting document, we get author name, categories and publication date. After that in term frequency calculation, we get highly frequent term in the corresponding document. We get title term as a high frequent term, as title has the most focused term in entire document. Then we consider the highest scored terms from Random Walk Term Algorithm. After all consideration, we get an integrated form as a metadata.

3.7 To Make Queries

There are 14 different queries asking 10 different student of Khulna University to issue 3 queries each, according to TREC specification [1], and chose one best query for each document. We assign unique id number of each query that is characterized by the .QID tag.

3.8 To Perform Relevant Judgements

Relevant and irrelevant judgment is very difficult, as it varies from person to person. In our corpus, we calculate doc-

ument to document relevant judgment using metadata; here the primary key is categories.

For query based relevant judgment, 14 initial queries are used in Random Walk based retrieval system. After that we make a simple interface and ask 10 different IT students for judge 43 relevant and irrelevant documents.

3.9 To Claculate Precision and Recall

Precission and recall are calculated using the equation (4) and (5) respectively.

3.10 To Draw Zipf's Curve

The Zipf's curve is drawn.

4 Experimental Result and Analysis.

In our corpus there are 69 text documents, 14 different queries with 43 relevant judgments. These total text documents are being collected from the 'Prothom-alo', 'Kaler-Kontho' and 'Amader Desh' Bangladeshi news paper [14][15][16]. The top ten most frequent terms in our corpus are given in the 1.

TABLE 5.1

TERMS FREQUENCY ON OUR CORPUS (WITHOUT STOP WORDS).

Word	Frequency	Word	Frequency
জন্য	0.613%	ঢাকা	0.353%
টিপাইমুখ	0.340%	বাংলাদেশ	0.340%
দেশের	0.323%	বাঁধ	0.313%
বিলবোর্ড	0.303%	কথা	0.286%
তাঁর	0.273%	সালে	0.266%

TABLE 5.2

PRECISION AND RECALL OF QUERIES ON OUR CORPUS.

Query ID	Precision (%)	Recall (%)
.QID 3	57	66
.QID 6	57	80
.QID 7	71	83
.QID 11	57	80
.QID 14	57	100
Average	59.8	81.8

Zipf's curve of our proposed corpus is shown in Fig. 1. Experimental results confirm that the law is correct for small corpora. For large corpora, however, it is found that at about rank 5000 the curve dropped below that of Zipf's straight line. We could say that our corpus has good text collection and the analysis about term frequency is also right [2]. We also calculate precision and recall for queries, .QID 3, .QID 6, .QID 7, .QID 11 and .QID 14 (given in Fig. 2). We get average precision = 59.8% and average recall = 81.8% (see Table 2). As we are unable to commpare our results with any existing Bangla text corpus, we try to get a result according to recognized standard. Our precision and recall shows better results compared to standard values [12].

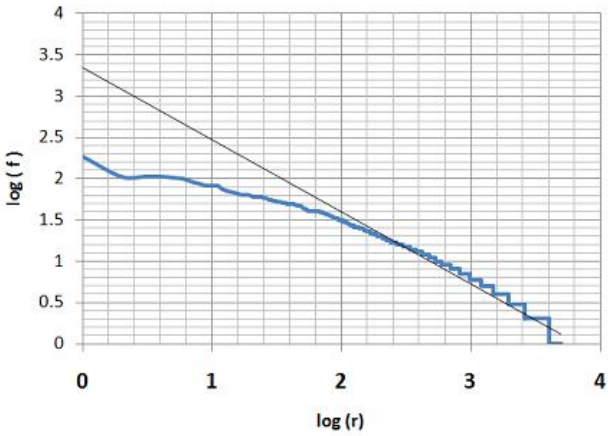

Fig. 1. Zipf's curve on Bangla corpus for 5000th rank

.QID 3

টিপাইমুখ বাঁধ এর পরিকল্পনা

.QID 6

বাংলাদেশের অবৈধ বিলবোর্ড এর সমস্যা ও করণীয়

.QID 7

শ্রীলঙ্কার জেনারেল ফনসেকার বিচার ও তার অপরাধ

.QID 11

বাংলাদেশে নৌ ও শ্রমিক ধর্মঘট

.QID 14

বাংলাদেশ ও ভারতের মধ্যে তিস্তার পানি চুক্তি

Fig. 2. Sample Queries

5 Conclusion

This is the first Bangla text corpus for IR system. We have proposed and implemented a novel method for building Bangla text corpus for IR purposes. We consider several criteria for evaluating a corpus for IR system, like priority based term frequency etc. Random walk on graph algorithm and making metadata to developing corpus are also used. Precision and recall shows the effectiveness of our corpus for IR.

6 Future Work

For Bangla text, there are number of complex sentence available and hence steaming is too much difficult to handle. In text collection for English text, text streaming is relatively easier and so it can be implemented easily. In this paper, we only consider few varieties of streamed text. A better result can be obtained by using a complete streaming

technique. This work can be extended by developing a good streaming algorithm for Bangla text.

ACKNOWLEDGMENT

We are very much thankfull to Dr. Md. Rafiqul Islam for his initial idea.

REFERENCES

[1] A. AleAhmad, H. Amiri and E. Darrudi. "Hamshahri: A Standard Persian Text Collection", Research report, school of electrical and computer engineering, university of Tehran.

[2] A. Bharati, R. Sangal, S. M. Bendre., 17-19 Dec. 1998, "Some Observations Regarding Corpora of some Indian Languages". Proc. Intl. Conf. Knowledge Based Computer Systems (KBCS- 98), NCST, Mumbai.

[3] Abu Shamim Md. Arif, M. M. Rahman and S. Y. Mukta., ICCIT 2009, "Information retrieval by modified term weighting method using random walk model ith query term positioning ranking". ICCDA, Paper ID: D223.

[4] B. B. Choudhury and U. Pal., 13th May, 1996, "A complete printed Bangla OCR system", Computer vision and pattern recognition unit, Indian Statistical Institute, Calcutta.

[5] D. C. Blair and M. E. Maron, "An Evaluation of Retrieval Effectiveness for Full-Text Document Retrieval System", working paper no:-364, Division of research, Graduate School of Business Administration, The University of Michigan.

[6] D. Hiemstra and D. van Leeuwen, "Creating a Dutch information retrieval test corpus" University of Twente, CTIT.

[7] H. Kucera and W. N. Francis "Brown corpus", Computational Analysis of Present-Day American English (1967).

[8] K. Md. Y. A. Majumder, Md. Z. Islam, and M. Khan, "Analysis of and Observations from a Bangla News Corpus", Center for Research on Bangla Language Processing, BRAC University, Dhaka, Bangladesh.

[9] M. R. Islam and S. A. Rajon., ICCIT 2009, "On the design of an Effective Corpus Evaluation of Bangali Text Compression Criteria", Khulna, Bangladesh.

[10] Md. T. Islam and S. M. Al. Masum "Bhasa: A Corpus-Based Information Retrieval and Summarizer for Bengali Text".

[11] N. S. Das, 2005, "Corpus linguistics and Language Technology", pp 93-219.

[12] R. P. Futrelle X. Zhang, Biological knowledge laboratory and scientific database project. College of Computer Science, 161 Cullinane Hall, Northeastern University, Boston, M. A - 02115. futrelle, xzhang}@ccs.neu.edu.

[13] http://members.unine.ch/jacques.savoy/clef/index.html

[14] Daily newspaper kalerkantho from http://www.dailykalerkantho.com

[15] Daily newspaper prothom-alo from http://www.prothom-alo.com

[16] Daily newspaper amaderdesh from http://www.amardeshonline.com

Jubayer Shamsed is an undergraduate student of Computer Science and Engineering (CSE) Discipline, Khulna University, Bangladesh. He has started his B.Sc.Engg.(CSE) degree in 2005. He is currently in his final year and doing his undergraduate thesis in the field of information retrieval on text corpus. He has particularly shown his keen interest in Bangla text corpus building. He is also a co-author of an international conference paper.

S. M. Masud Karim has been serving as a faculty member of Computer Science and Engineering (CSE) Discipline, Khulna University, Khulna, Bangladesh. He completed his B.Sc.Engg.(CSE) degree with distinction in 2001. He went abraod for hisher studies in 2006 and was awarded M.Sc. in Media Informatics from Technical University of Aachen (RWTH Aachen), Germany in 2008 and M.Sc. in Informatics from University of Edinburgh, UK in 2009. He has a good number of international research papers and has supervised a good number of undergraduate thesis students. His areas of interest include information retrieval, data exchange, data integration, and computer security.