Construction of Durian Dataset from Web Collection for Query Reformulation Research

Azilawati Azizan, Zainab Abu Bakar, Nurazzah Abdul Rahman

Abstract— ***Study in the field of Information Retrieval (IR) has long been developed and thrived over time. And most of them use the available standard dataset for testing and evaluation. In line with that, the existence of new dataset has also increased to meet the needs of their respective studies. However, to the best of our knowledge, there is no dataset collected from web document that focuses on fruit domain. Therefore, in this paper we contribute to this field by publishing a dataset of web document for fruit focusing on durian fruit. This durian fruit dataset is suitable for query reformulation experiment, searching system, web information retrieval and any search engine experiment. This dataset contains a collection of web document for fruit and durian fruit, a collection of queries and a set of relevant judgement. In addition, in this paper we also publish a list of frequently asked query regarding durian, and an extended list of query characteristic categories.***

Keywords: Dataset construction, durian, relevant judgement, test query, web collection.

I. INTRODUCTION

Dataset is used to measure the effectiveness of the Information Retrieval (IR) system. It measures how well the IR systems perform, which normally compare the performance of the IR system with other IR system, or compare search algorithm with other algorithm and compare search strategies. A complete IR dataset commonly consist of document set, query set and relevant judgement set. Document set is a collection of all documents related to the chosen scope. While query set is a set of information needs which is represented in collection of questions asking the IR system for results, and relevant judgement set is a set of known relevant documents for each of the information needs (the question). Currently there are many datasets available [1] and most were originally developed to support research in IR, but practitioners often find them useful as well.

In conducting our research about query reformulation for finding domain specific information on Web, we have constructed the dataset beforehand in order to test and evaluate the techniques that we proposed. We constructed the durian dataset due to our predefine research’s scope. We do not use any standard IR dataset because there is no durian dataset available for use during the research.

Durian belongs to the category of fruits and the larger category is agriculture. From our observation and searches made, we found out dataset on agriculture and fruit are very limited compared to dataset from other field such as news, medical and general. Coincidence with that, we feel it is necessary for us to publish this dataset for public use. In this paper, we describe all steps taken in constructing the dataset. We have constructed a complete dataset which consist of durian related document set, durian query set and relevant judgement set. This dataset can be used in any IR system for testing and evaluation, specifically focused for durian and fruit domain.

II. RELATED WORKS

A. Test data

Currently, there are many datasets available online for research and academic use. Different field of research needs different types of datasets [2]. Among the popular and standard dataset for IR field are the Text Retrieval Conference (TREC), Cranfield collection, Cross Language Evaluation Forum (CLEF), NII Test Collection for IR system (NTCIR) and others. Each of those datasets has their own target area in the information retrieval field [3].

The pioneer test collection in IR that allows precise quantitative measure is the Cranfield collection [4]. However, in these days the collection is too small to be used for testing, but it has been the most elementary pilot experiments in IR.

On the other hand, the most popular dataset in IR is TREC. TREC is co-sponsored by the National Institute of Standard and Technology (NIST) US Gov. and the Information Technology office of Defense Advanced Research Project Agency (DARPA). This collection started since 1992. It has several different tracks and topic to cater different research areas, and TREC-CLIR (Cross-Lingual Information Retrieval) track provide for multilingual evaluation [5].

CLEF is another dataset commonly being used in earlier IR research. It was founded by the European Union in 2000. It provides test collection/data for multilingual information retrieval. The CLEF focus for cross language in European languages, and it also supports multilingual retrieval for multimedia evaluation.

NTCIR (NII Test Collection for Information Retrieval system) provides a large-scale reusable data set test collection but its focus more specifically for languages form East Asian such as Japanese, Chinese and Korean. It is coordinated by National Institute of Informatics (NII) in Japan.

Revised Version Manuscript Received on September 16, 2019.

Azilawati Azizan, Faculty of Computer and Mathematical Sciences, Universiti Teknologi MARA, Perak Branch, Tapah Campus, 35400 Tapah Road, Perak, Malaysia.

Zainab Abu Bakar, Al-Madinah International University, 40100 Shah Alam, Selangor, Malaysia.

Nurazzah Abd Rahman, Faculty of Computer and Mathematical Sciences, Universiti Teknologi MARA, 40450 Shah Alam, Selangor Malaysia.

III. MOTIVATION

Agriculture is a big domain consist of several sub areas. There are three main areas in agriculture at Malaysia. It is divided into crops, fishery and livestock. Among the areas, crops have the most sub areas in it, such as industrial crops, paddy, fruits, vegetables, cash crop, spices, herbs and floriculture. Agriculture continually been given more attention in our country (Malaysia). In rapidly developing countries such as Japan have also never abandoned this sector in enhancing its development. In Malaysia, agriculture is one of the areas that significantly helps to raise the country's economy. Due to that, the government arrange many initiatives to the farmers for them to increase the productivity and indirectly raise the national economy.

Former Prime Minister of Malaysia, Dato' Seri Mohd. Najib Tun Abdul Razak in his presentation of the national budget for 2014, has stated that the agricultural industry has always been the important agenda. This is proved by the government by allocating RM6 billion for the implementation of agricultural programs that have high added value and commercial value [6].

Many local farmers in Malaysia are still far left behind in finding related information using search technology on the Web. Many of them are still depending on phone calls, face to face discussion with expert and book / manual reference. This statement is proved by the expert in agriculture from MARDI (Malaysian Agricultural Research and Development Institute). This is due to the lack of applications that can help them to improve their task. So, there is a need to have a retrieval application that can benefit them with their task align with the current technology trend.

Durian is chosen to be the domain area in our research because 96% of Malaysian loves to eat durian and 60% Malaysian prefer buying durian compared to other types of fruit. This shows durian have a good market and have high potential for improvement [7]. In particular, the former Prime Minister also said that the country will increase the export of durian fruit to foreign markets, particularly China, as demand and high prices for the durian fruits there [6]. Other than that, durian is also a significant factor for maintaining good family relationship, since 63% of Malaysian still return to hometown just to eat durian [7].

Furthermore, durian also is a popular search term for tropical fruit. It is based on Google Trend results (Fig. 1) showing for ten years of trends starting from 2009 to 2019. As in Fig. 1, durian search term (the uppermost graph line at Fig. 1) is keep increasing from years to years. Moreover, durian is also getting the highest trend compared to other tropical fruit such as *rambutan* or mango.

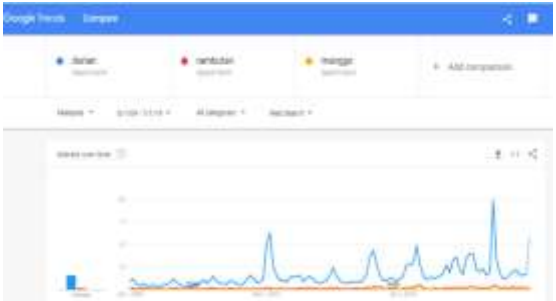

Fig. 1: Google trend for "durian" search term

Fig. 2 shows that Malaysia has the most interest on finding about durian on Web. Therefore, it is relevant enough for us to publish this durian and fruit dataset.

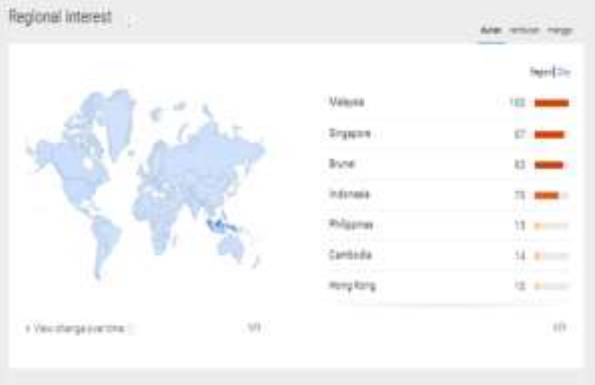

Fig. 2: Regional interest of the "durian" search term

IV. METHODOLOGY

A. Document set

The document set were collected from the Web, which is in HTML format. A crawler was used to collect all related HTML documents on the Web. The crawler crawl over the Web using provided keyword. The keywords being fed to the crawler were 'durian' and 'durio'. Durio is another term used for durian. Fig. 3 shows the illustration of the methodology used in constructing the dataset.

In the initial stage, 200 HTML documents were gathered. Then in order to further strengthen the test collection, 3000 HTML documents were gathered later. On the other hand, keyword 'fruit' was also been fed to the crawler to gather the non-related documents, which contain other fruits documents such as rambutan, strawberry, banana and others.

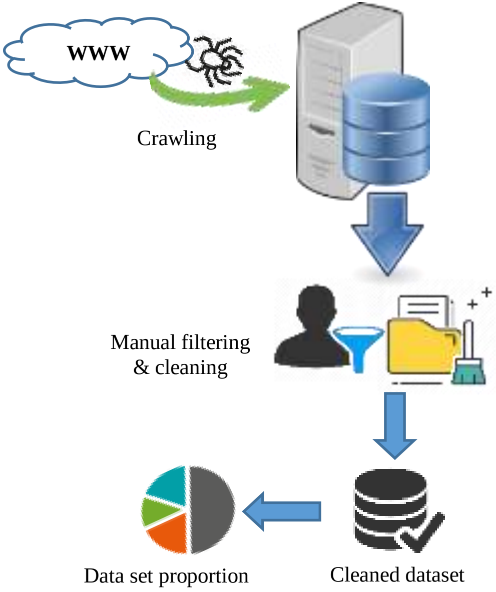

Fig. 3: Methodology of the dataset construction

B. Query set

Various sources have been used to gather the queries. The queries were collected from two types of user; the general users and the domain expert users. General user comprises of Web and non-Web users. The Web users were from the farmers' online forum, blogs, social networks (Facebook, Tweeter) and the Google suggestion system features (Google Instant). Second source of the query is from the domain experts that are the durian expert from MARDI and the durian farmers. Fig. 4 illustrates the sources being used to gather the query set.

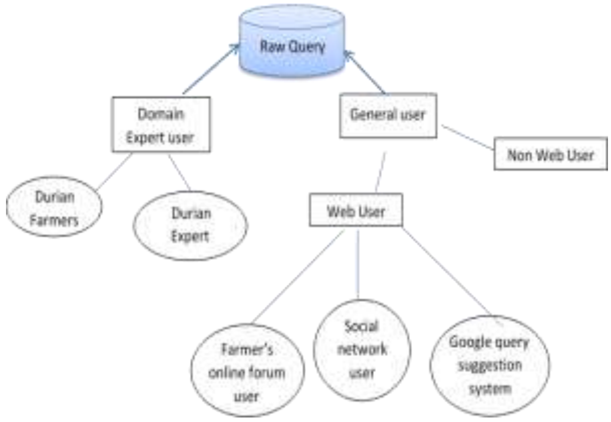

Fig. 4: Query gathering sources

MARDI is a statutory body which has been mandated to conduct research in agriculture, food and agro-based industries in Malaysia. General characteristic for crops has been determined by MARDI on several discussions and on few books issued by them. There are ten general characteristics for crops as shown in Table 1

Table 1: Crops general characteristic by MARDI

N o.	Category
1	Description
2	Varieties
3	Uses
4	Propagation
5	Culture and Management Practice
6	Pests
7	Disease
8	Fertiliser
9	Harvest
1 0	Post-harvest

Then, in [8] has expanded the list to 21 categories which based on the study of facts about durian on Web documents. However, in our study, when the entire query being analyzed the categories has been expanded to 25 categories. 4 new categories (location, supplier, picture of durian and others) were added to suit with the query collection. It was added based on the queries group created in the process. See Table 2.

Table 2: Durian characteristic category

No.	Category
1	Description
2	Varieties
3	Uses
4	Propagation
5	Culture and Management Practice
6	Pests
7	Disease
8	Fertiliser
9	Harvest
10	Post-harvest
11	Belief
12	Nutrition
13	Health
14	Odor and Flavor
15	History
16	Market
17	Season
18	Price
19	Serving fresh fruits
20	Recipe
21	Growing stage
22	Location
23	Supplier
24	Picture of durian
25	Others

Facts about durian were gathered from the MARDI experts. A discussion and interview session have been set up to acquire more facts and to verify all related information and sources with the experts. Besides that, Internet sources were also been used to enrich the facts which has been verified by MARDI at the end of the process. Below is the list of all sources that has been used.

- Domain Experts–MARDI (Malaysian Agriculture Research and Development Institute)
- Internet Sources–general information
- Internet information sources are from trusted website.
- iTFNet–International Tropical Fruits Networks. (http://www.itfnet.org/)
- MOA-Ministry of Agriculture (http://www.moa.gov.my/)
- DOA-Department of Agriculture (http://www.doa.gov.my/) website
- MARDI-Malaysian Agriculture Research and Development Institute (http://www.mardi.gov.my/home)
- Jabatan pertanian negeri Perak–(http://www.pertanianperak.gov.my)
- Jabatan pertanian negeri Penang–(http://jpn.penang.gov.my/#)

C. Relevant judgements

Collaboration with MARDI has made the relevant judgment process more reliable. Discussion with durian experts from MARDI has given a lot of knowledge and information about durian fruit.

All selected queries were listed properly in a document in advance. 30 queries have been selected to be used in the experiment and retrieval testing. Then, a discussion with experts from MARDI was organized. All the procedures and criteria for constructing the relevant judgement were laid out with 3 main experts involved in the collaboration. Each documents were checked manually in order to identify and labeled whether the documents is relevant or not relevant for the respected queries.

V. RESULTS AND DISCUSSION

There were 3000 documents (web documents) gathered for the non-related durian documents and 3000 documents (web documents) for durian related documents, which have in total of 6000 documents for the document set. This means, the dataset has 50% related and 50% non-related documents to the domain [9]. Table 3 shows the list of all documents gathered with the percentage which contribute to the document collection.

Table 3: Table of document percentage in the dataset

Fruit Name	Percent age
Durian	50
Fruit-Mix	15
Banana	14
Strawberry	7
Avocado	3
Mango	2
Mangosteen	1
Carambola/Star Fruit	1
Soursop	1
Guava	1
Rambutan	1
Grapes	1
Pineapple	1
Pitaya	1
Jackfruit	1

The proportion of the dataset is illustrated in pie chart as in Fig. 4. As stated in previous section, the dataset consist of 50% document about the durian, and the other half consist of various type of fruit document. The selection of other fruits is based on the most trending search terms from Google Trend analysis results.

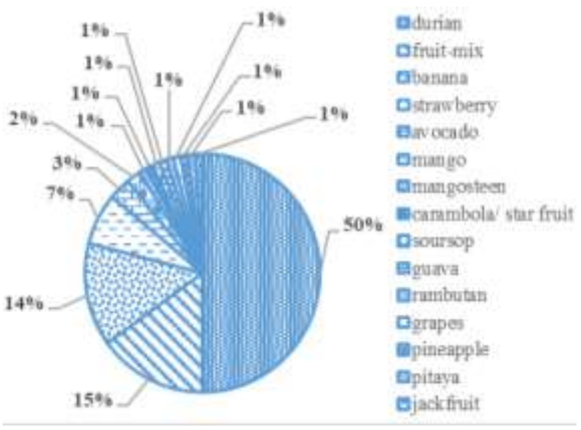

Fig. 4: Pie chart of the dataset proportion

A. Test query-most common query

Expert from MARDI have marked all the queries either less frequent, average and most frequent being asked by user. Then for this study, only query marked with average and most frequent were selected to be used in the testing.

Finally, 30 queries were selected to be the test query. Based on [10], the author has selected 21 query from forums to be their test query, which gave a baseline for us to use 30 number of test query. Table 4 shows the test queries for this research. These 30 queries have been verified by MARDI expert as the most frequently asked queries by user about durian.

Table 4: Test query- most frequent query about durian

Query No.	Test Query
Q1	List of insect pests that attack the durian tree.
Q2	When is the durian season in Malaysia?
Q3	What are the varieties of durian in Malaysia?
Q4	What are the characteristic of good quality durian?
Q5	How to plant durian?
Q6	How to control durian tree disease?
Q7	What are the products of durian?
Q8	What are the side effects of eating durian to health?
Q9	Types of land suitable for durian tree planting.
Q10	What are the durian tree diseases?
Q11	How durian tree disease can be treated?
Q12	What is the best durian clone?
Q13	Why durian tree does not fruitful?
Q14	How to make durian tree fruitful quickly?
Q15	Finding durian supplier.
Q16	Why some people do not like to eat durian?
Q17	Where to get durian?
Q18	What are the durian tree plantation stages?
Q19	What are the insect pests that attack durian tree?
Q20	Information about water irrigation for durian plantation.
Q21	Suitable fertilizer for durian tree.
Q22	The most delicious durian type in Malaysia.
Q23	Types of durian tree.
Q24	The best durian seedlings.
Q25	Where to get durian seedling?
Q26	How to choose a good durian?
Q27	Nutrition facts of durian fruit
Q28	Durian fruit specialty.
Q29	Ways to pick a good durian.
Q30	Why durian is so tasty.

Finally, we summaries our Durian dataset details as in Table 5. There are 6000 documents in total, with 319 raw queries, 30 test queries which has been labeled as the most frequent queries being asked by user about durian and with a set of relevant judgement for all the 30 test queries. Our dataset is in flat file (txt) format which has about 478MB of file size.

Table 5: Information of durian dataset

Item	Information
Document Set:	
Stemming	No
Webpage (HTML)	6000
Preprocessed (cleaned)	Yes
Query Set:	
Raw query	319
Distinct query (cleaned)	105
Most common query	30
Query category	25
Relevant Judgement:	
Relevant documents	Yes
Others:	
Dataset file format	Flat file
Crawled Between	Jan-Apr 2013
Keyword Seed	Durian, durio, rambutan, fruit
Language	English
Limitation	No image
Dataset file size	478MB

VI. CONCLUSION

Main contribution of this work is the complete dataset containing document set, query set and relevant judgement of durian fruits. It is specifically design to be used in query reformulation research which also can be used in any IR system research and experiment. Along with this dataset construction, a list of most common queries being ask about durian also being laid out. Then, an extended list of durian characteristic category also has been figure out for public use. Other than that, this paper also has shared the methodology used for constructing the dataset which has been collected from the Web for the document set. We hope with the existence of this Durian dataset, will benefit the researcher who wish to conduct research in IR, specifically those relating to fruit domain study. Besides, this work also contributes to the agriculture field, and hopefully it will greatly benefit the researchers and farmers from Malaysia.

VII. ACKNOWLEDGMENT

The research is supported by Malaysian Government and Universiti Teknologi MARA (UiTM), specifically the Faculty of Computer and Mathematical Sciences (FSKM), Universiti Teknologi MARA, Perak Branch. The authors are indebted to the reviewers for their valuable comments and suggestions, which led to the improvement of the paper.

REFERENCES

1. A. Alakrot, L. Murray, and N. S. Nikolov, “Dataset construction for the detection of anti-social behaviour in online communication in Arabic,” Procedia Comput. Sci., 142, 2018, pp. 174–181.
2. Y. Yao, J. Zhang, F. Shen, X. Hua, J. Xu, and Z. Tang, “A new web-supervised method for image dataset constructions,” Neurocomputing, 236, 2017, pp. 23–31.
3. B. S. J. Brenda, Standard datasets in IR. Available: https://www.slideshare.net/JeeeeNU/sds-ir.
4. C. W. Cleverdon, “The significance of the cranfield tests on index languages,” 14th Annual International ACM SIGIR Conference on Research and Develpoment in Information Retrieval, 1991, pp. 3–12.
5. NIST, Text REtrieval Conference (TREC) Home Page. Available: http://trec.nist.gov/.
6. BERNAMA, RM6b pelaksanaan program pertanian. Available: htttp://mynewshub.my/2013/10/25/rm6-bilion-untuk-pelaksanaan-program-pertanian/#ixzz2jkkqPeFd.
7. R. A. Dardak, “Kecenderungan makan durian di kalangan penduduk Malaysia,” Econ. Technol. Manag. Rev., 1, 2006, pp. 37–49.
8. Z. A. Bakar, K. N. Ismail, and N. Fooladi, “Effectiveness of query formulation based on durian characteristics,” IEEE Student Conference on Research and Development, 2012, pp: 59-62.
9. T. Sabbah, A. Selamat, M. H. Selamat, R. Ibrahim, and H. Fujita, "Hybridized term-weighting method for dark web classification," Neurocomputing, 173, 2016, pp: 1908-1926.
10. M. Radja, B. S. Nathalie, S. Michel, and M. Mireille, “Ontology-based reformulation of health consumer queries using spreading activation techniques,” Int. Conf. Intell. Comput. Cogn. Informatics, 2010, pp. 98–101.