Detecting Cropping Patterns of Underutilized Crops using Online Big Data

Ayman Salama Mohamed
CropBASE Program,
Crops For the Future (CFF),
Selangor, Malaysia
ayman.mohamed@cffresearch.org

Ebrahim Jahanshiri
CropBASE Program,
Crops For the Future (CFF),
Selangor, Malaysia
ebrahim.jahanshiri@cffresearch.org

Tomas Maul
School of Computer Science,
University of Nottingham Malaysia Campus,
Selangor, Malaysia
Tomas.Maul@nottingham.edu.my

***Abstract*—Crop that are currently underutilized can play a major role in diversifying food sources and combating climate variability. One major obstacle for wider adoption of these species is the lack of information on the geographic areas where these crops are currently grown. These crops are typically grown in marginal lands through subsistence agriculture. At present, there is no global database and no efficient procedure that allows users to acquire cropping patterns of underutilised crops. The proposed solution identifies underutilized cropping patterns using online search engine data. The target is to determine global public interest in underutilised crops over time through search engine data. This identifies possible crop utilisation patterns, trends and interest pertaining to underutilised crops over time. As a proof of concept, we collected a set of keyword synonymous of Bambara groundnut (BG), from local and international databases and research publications. Using the Google AdWords service and 40 different terms for BG, we were able to gather search event data for two years, at the city level. Preliminary analysis done through a software prototype shows that the data provides new insights as to how BG search events are distributed and how this data can be used to delineate current areas where BG grows and what are the characteristics of their value chains. For evaluation purposes, we compared our BG results with the crop's known network of BG growers and researchers, and confirmed that the results not only matched known regions of the network but also proposed several new ones that need to be evaluated. The results suggest that the proposed solution provides significant indicators for possible cropping patterns and/or research interests around the world.**

Keywords—Underutilised crops; data mining; Google keywords; cropping patterns

I. INTRODUCTION

Given the importance of agricultural biodiversity, species that are currently underutilised, can play a major role in diversifying food sources and combating climate variability [4]. Recent research into biodiversity is significantly increasing and contributing towards the development of a confined data repository and knowledge base [1]. Establishing a database of the current regions and localities where these crops are grown can help scientists, farmers and other decision makers, to rapidly establish a global network for information exchange. The established network will in turn help with the data collection of local agricultural practices and value chains for these crops that will streamline wider adoption and inclusion in agricultural diversification plans worldwide [23].

The current advancement of research on and adoption of, underutilised crops is crucial. Significantly it can help in mitigating the effects of global problems like climate change, nutrition deficiency, food security and to decrease developing countries' dependency on imported crops [7]. The research on underutilised crops will contribute to the achievement of the United Nations Sustainable Development Goals (No Poverty, Zero Hunger and Climate Action) [9]. Data on major crops like rice, wheat, maize and soybean are significantly available given the fact that they have their own dedicated research centres such as IRRI for rice [21] and CIMMYT for maize [16].

Underutilised crops suffer from a lack of dedicated research centres which leads to unavailability of centralized and organized data. Research on underutilised crops requires identification of the places where the crops are grown, planted and used as food, feed or in commercialized products. This will lead to the detection of existing value chains and communities that are involved. The first goal of this research involved identifying the various naming conventions used for the target crop. For this purpose, an experiment was conducted to collect and verify all possible names for the crop before extracting search engine data based on those keywords.

Every crop can have many varieties, e.g.: the apple fruit has more than 7500 varieties [7]. The same varieties of apple can be referred to by different names in different languages. Moreover, it can be called different names in the same language based on geographical locations. Varieties, local names, synonyms and landrace names are different names that can define certain crops [24]. For major crops like rice and maize, most of the names have been documented. Most of their varieties have their genetic sequences stored in gene banks [19]. This is not the case for most underutilised crops. Their varieties, landrace and cultivar names and their cropping patterns are not well documented. Solving this problem requires an expert study on underutilised crop names. This study constitutes a preliminary step in this research. Web crawling and data mining approaches are particularly useful in ascertaining social trends [22], and have also proven to be useful in biological and agricultural sciences. Van der Velde et al. (2012) have used search volume data to generate crop specific planting and harvesting information, integrating climate and soil data [23]. Search engine data has also been used in health care to predict trends and disease hotspots [2], [10].

This research is focused on identification of possible cropping patterns of underutilized species using online search engine data. The target is to collect and analyse search engine data that can be categorized per location and over time. This will identify the possible crop utilization patterns as well as trends and interest on underutilised crops over time.

II. Bambara groundnut Data

A. Choice of Bambara groundnut

Crops For the Future (CFF) is a dedicated research centre for underutilised crops. Bambara groundnut is one of the main crops that CFF is researching [14]. CFF established the BamYield network to link various research centres and institutions that are working on Bambara groundnut. This research is using Bambara groundnut as an example of using the Google AdWords tool to identify cropping patterns by investigating global public interest through search engine requests of various Bambara groundnut names. The existing data from the CFF BamYield program will help to validate the results of the research.

B. Bambara groundnut Names

The initial stage of this research was to identify Bambara groundnut names, scientific names, synonyms, varieties, landraces, cultivars and local names in different languages. The process of investigating Bambara groundnut names involved several experts, extracting names from the literature and integrating them with local and international databases. The international databases that were used in this research consisted of United States Department of Agriculture, Germplasm Resources Information Network USDA-GRIN [15], Global Biodiversity Information Facility GBIF [20], Food and Agriculture Organization AGROVOC [3], and CFF CropBASE [12], [13]. The initial list of Bambara groundnut names that had been composed from international databases and the literature contained several incorrect names. The reason for these inaccuracies stems from the limited interest in those crops. A manual process was employed to filter and enhance this list. Table 1 lists the verified Bambara groundnut names.

TABLE I. Bambara groundnut Names

Bambara groundnut names	
Group 1	***Group 2***
Bambara groundnut	kacang bogor
Bambara Bean	ถั่วหรั่ง
Bambara-bean	バンバラマメ
Bambara Erdnuss	Feijão jugo
Bambara-Erdnuss	班巴拉花生
Bambarra groundnut	Congo goober [6]
Guisante bambarra	Erderbse
Madagascar groundnut	jinguba-de-Cabambe
Jugo bean	비그나서브테라네아
vigna podzemní	Pisello di terra
Voandzeia subterranean	Pois arachide
Vigna subterranean	Voandzou
Glycine subterranean	kacang poi
Maní de bambarra	бамбарский земляной орех

III. Method

Every day billions of search requests are performed on search engines from all over the globe. Recently several research efforts from various disciplines used search engine data to extract valuable information and indicators [18]. These keyword search logs and statistics, which are essentially crowd data, are normally provided by the search engines in different formats.

A. Google Keywords Tool

Google processes over 3.5 billion requests per day [11]. Given the multitude of Google and its collaborative search engines users, the search event data from these databases can be used to demonstrate public interest in particular subjects [2], [10], [23]. The Google AdWords application is used by companies to allow customers to monitor their business. Google AdWords provides several planning tools. Keyword planner [8] for example is used to decide which keyword to link to customer business. One function of this tool is to get the current trend on any desired keyword. Keyword planner takes certain words or phrases as input and the output is the number of search requests that happened in certain locations and certain periods of time.

B. Google Trends Tool

The Google Trends tool has been studied in this research. It provides detailed trends on specific search keywords [5]. Several research efforts in various disciplines started to use Google trends as a tool to investigate social interactions with their experiments [22]. However, this tool has a limitation which consists of displaying trends as percentages rather than raw counts. Google Trends takes the highest search result count in a specific time or place and represents it as 100%. All other search results will be percentages relative to this maximum result. In other words, Google Trends never displays the actual search request raw counts. Google Trends also has a minimum threshold for search request counts in order to display the trends. If the counts are less than this threshold, no data will be displayed. Since underutilised crop searches tend to involve significantly small counts, as will be shown in the results section, Google Trends can't be used as the main tool of this research. However, it was used to verify Google keyword results, and it was also used to get extra keywords that appear in the recommendation list of the tool.

C. Using Google Keywords Tool

This section describes in detail the steps taken and configurations used in the Google Keyword planner tool to extract counts of search requests for specific keywords.

- Input. The Google Keywords application takes specific keywords in text or csv format as input. The list of 40 Bambara groundnut names that were stated above were uploaded as a csv list.
- Configuration. Target output: volume data. Search engine choice: Google search engine and its partners. Geographical location: all locations. Chosen period of time: from July 2014 to June 2016. The tool has to be run manually for each country.

- Output. The tool provides a list of search request counts per country and per month. The tool breaks down the results to show the count per device that was used for searching (i.e. desktop, mobile or tablet).

IV. Results and Data Analysis

The extracted data consists of 230 csv files. Each file contains a list of search request counts in the specified country, per month, per device, per keyword. The 230 files were processed using shell scripting on Linux Centos to extract cropping patterns.

A. Top Countries that Search BG

A geographical web interface was constructed to present the data geographically using Google GeoChart. The map in Fig. 1 shows the average search request counts for Bambara groundnut names per country for the period of (July 2014 – June 2016). Top countries are (Indonesia, 1110), (Thailand, 470), (United States, 400), (South Africa, 320), (Malaysia, 320), (Nigeria, 260) and (United Kingdom, 220).

B. Local Names and Langagues for BG

Bambara groundnut in Indonesia is a commercialized crop. Customers can buy its products from the market [17]. The local name for Bambara groundnut in Indonesia is "kacang bogor". This name is the highest searchable keyword from the list of Bambara groundnut names. Kacang bogor takes 91% of the search requests in Indonesia from all Bambara groundnut names. Thailand is the second in the list. The most searchable name on Google for Bambara groundnut names is "**ถั่วหรั่ง**". The United States of America is the third country in the list. We assumed that this was mostly due to research institution's search requests and expected accordingly that the top searchable name was *Vigna* subterranean, which is the scientific name of Bambara groundnut. Surprisingly the top two names that are used in the United States in search requests are Bambara groundnut and Bambarra groundnut (with double "r"). Preliminary investigation shows that African Americans are starting to reconnect with their ancestors' crops, including Bambara groundnut. Both the United Kingdom's and South Africa's top searchable name is Bambara groundnut. Malaysia's top searchable name is the local name: "Kacang poi".

C. Temporal Analysis of Google Search Requests Data

Time is the second parameter that the Google Keywords tool generates. Temporal analysis provides the ability to show search trends over time. The graph in Fig. 2 indicates the trend of search requests in a period of two years (July 2014-June 2016). The blue line shows the Indonesian trend. There are two main peaks in March 2015 and March 2016. After investigating those two peaks with Bambara groundnut experts, they confirmed that those peaks happen around harvest time each year. The Indonesian search trends show a significant increase of interest from 2015 to 2016. This probably means that there is an increase of public interest in Bambara groundnut and its products overtime. Similar trends have been observed in Thailand around November 2015 and October 2016, however we were not able confirm whether this corresponds to harvest time in Thailand.

D. Search Request per Device

The third element in the tool's set of results consists of the device used to initiate search requests. The graph in Fig. 3 shows the average number of search requests per month, per device (Desktop/Mobile). The primary assumption is that a higher number of desktop-based searches might be linked to research institutions, while a higher number of mobile and tablet search requests might indicate high usage from the common public and farmers. Indonesia, Thailand and Malaysia show higher numbers of search requests from mobile devices which suggests there is significant interest by the common public.

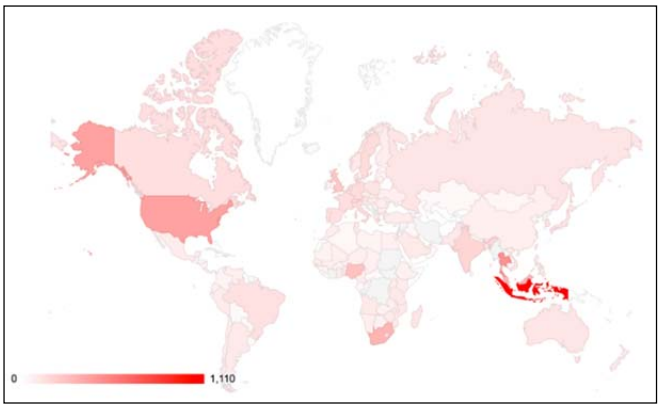

Fig. 1. Average Google search requests for BG.

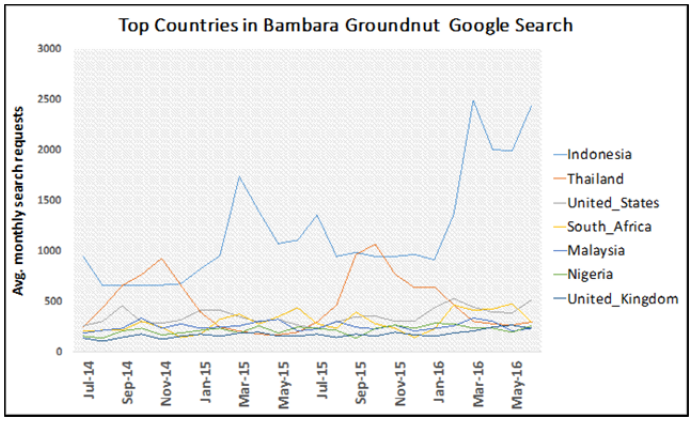

Fig. 2. Top countries for Bambara groundnut Google search.

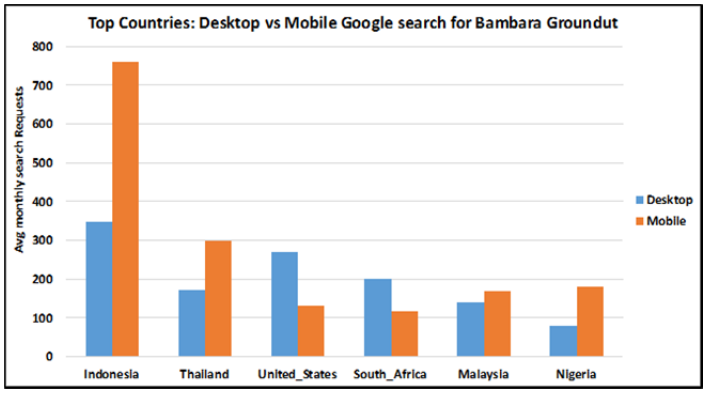

Fig. 3. Top countries for Bambara groundnut Google searches (Desktop vs mobile).

These results are consistent with the fact that Bambara groundnut is commercialized in those countries and several products depend on it. However, the United States and South Africa have a higher number of search requests from desktops which suggests a higher interest from the research and industry sectors.

V. Discussion

The Google AdWords tool was proven to be a valuable source of information for the cropping patterns of underutilised crops. Google AdWords has a limitation when the search requests are below ten requests per month. Average search requests per month are depicted as multiples of ten, which can be interpreted as the tool's rounding mechanism. Thus, studying countries with small numbers of monthly search requests was ineffective because of the tool's rounding mechanism.

When using Google's search engine, users typically type phrases, not only keywords. This was not considered in this research due to the generally small number of search requests involved. However, the Google trends tool provides top phrases or similar keywords that people might use for certain topics. This option was studied in this research in order to get similar search keywords or phrases. However, this was shown to be ineffective because of the significantly small number of search requests for underutilised crops.

The United States of America appeared to be number 3 in the top list of search requests. This finding is new to the BamYield network and shall be considered in the coming analysis of Bambara groundnut countries to explore. In spite of Bambara groundnut originating from and currently being utilized in Africa, the search request counts from African countries are too small, which makes it impossible, at this stage, to extend the analysis and validation to this broader context. One possible extension of this research involves utilizing other social network data that might complement Google AdWords data. Future work will include more underutilized crops like Moringa (Moringa olifera) and Winged bean (Psophocarpus tetragonolobus), and might also include the validation of results using alternative crowdsourced data.

References

[1] Ananiadou, S., Batista-Navarro, R., & Zerva, C. "Construction of a Biodiversity Knowledge Repository using a Text Mining-based Framework." SIMBig. (2016).

[2] Bardak, Batuhan, and Mehmet Tan. "Prediction of influenza outbreaks by integrating Wikipedia article access logs and Google flu trend data." In Bioinformatics and Bioengineering (BIBE), 2015 IEEE 15th International Conference on, pp. 1-6. IEEE, 2015.

[3] Caracciolo, Caterina, Armando Stellato, Ahsan Morshed, Gudrun Johannsen, Sachit Rajbhandari, Yves Jaques, and Johannes Keizer. "The AGROVOC linked dataset." Semantic Web 4, no. 3 (2013): 341-348.

[4] Chivenge, Pauline, Tafadzwanashe Mabhaudhi, Albert T. Modi, and Paramu Mafongoya. "The potential role of neglected and underutilised crop species as future crops under water scarce conditions in Sub-Saharan Africa." International journal of environmental research and public health 12, no. 6 (2015): 5685-5711.

[5] Choi, Hyunyoung, and Hal Varian. "Predicting the present with Google Trends." Economic Record 88, no. s1 (2012): 2-9.

[6] Directorate Agricultural Information Services Department of Agriculture, Forestry and Fisheries Private Bag X144, Pretoria, 0001 South Africa. "Production guideline for Bambara groundnuts." (2011).

[7] Ebert, Andreas W. "Potential of underutilized traditional vegetables and legume crops to contribute to food and nutritional security, income and more sustainable production systems." Sustainability 6, no. 1 (2014): 319-335.

[8] Geddes, Brad. Advanced Google AdWords. John Wiley & Sons, 2014.

[9] Griggs, David, Mark Stafford-Smith, Owen Gaffney, Johan Rockström, Marcus C. Öhman, Priya Shyamsundar, Will Steffen, Gisbert Glaser, Norichika Kanie, and Ian Noble. "Policy: Sustainable development goals for people and planet." Nature 495, no. 7441 (2013): 305-307.

[10] Herman Anthony Carneiro1,2 and Eleftherios Mylonakis1, "Google Trends: A Web-Based Tool for Real-Time Surveillance of Disease Outbreaks". Clin Infect Dis. (2009) 49 (10): 1557-1564. doi: 10.1086/630200.

[11] Internet live stats. "Google Search Statistics" internetlivestats.com. http://www.internetlivestats.com/google-search-statistics (accessed March 13, 2017).

[12] Jahanshiri, E., Walker, S., Ehsan, S. D., & Jahanshiri, E. Paper Reference Number: 47 Collaborative Agricultural Research Environment Based on Open Source Geospatial Web Technologies.

[13] Jahanshiri, Ebrahim, and Sue Walker. "Agricultural Knowledge-Based Systems at the Age of Semantic Technologies." International Journal of Knowledge Engineering-IACSIT 1 (2015): 64-67.

[14] Mayes, S., F. J. Massawe, P. G. Alderson, J. A. Roberts, S. N. Azam-Ali, and M. Hermann. "The potential for underutilized crops to improve security of food production." Journal of experimental botany 63, no. 3 (2012): 1075-1079.

[15] NRCS, USDA. "The PLANTS Database (http://plants. usda. gov). National Plant Data Center, Baton Rouge." Accessed March 30 (2010): 2010.

[16] Pingali, Prabhu L. "CIMMYT 1999/2000 World maize facts and trends. Meeting world maize needs: Technological opportunities and priorities for the public sector." (2001).

[17] Plahar, W. "Marketing and processing of bambara groundnuts (West Africa). Final Technical Report." (2009).

[18] Polykalas, Spyros E., George N. Prezerakos, and Agisilaos Konidaris. "An Algorithm based on Google Trends' data for future prediction. Case study: German Elections." In Signal Processing and Information Technology (ISSPIT), 2013 IEEE International Symposium on, pp. 000069-000073. IEEE, 2013.

[19] Sherry, Stephen T., M-H. Ward, M. Kholodov, J. Baker, Lon Phan, Elizabeth M. Smigielski, and Karl Sirotkin. "dbSNP: the NCBI database of genetic variation." Nucleic acids research 29, no. 1 (2001): 308-311.

[20] Taxonomy, GBIF Backbone. "doi: 10.15468/39omei." (2016).

[21] The International Rice Research Institute (IRRI). http://irri.org/about-us/our-mission

[22] Trevisan, Filippo. "Search engines: From social science objects to academic inquiry tools." First Monday 19, no. 11 (2014).

[23] Van der Velde, Marijn, Linda See, Steffen Fritz, Frank GA Verheijen, Nikolay Khabarov, and Michael Obersteiner. "Generating crop calendars with Web search data." Environmental Research Letters 7, no. 2 (2012): 024022.

[24] Villa, Tania Carolina Camacho, Nigel Maxted, Maria Scholten, and Brian Ford-Lloyd. "Defining and identifying crop landraces." Plant genetic resources: characterization and utilization 3, no. 03 (2005): 373-384.