Building Vocabulary for Big Data Analytics

Lyublyana Turiy
Palmer School of Library and Information Science
Long Island University
Greenvale, NY
Lyublyana.Turiy@my.liu.edu

***Abstract*—The explosion in Big Data Analytics research provides a massive amount of software capabilities, publications, and conference proceedings making it difficult to sift through and inter-relate it all. A vast amount of new terminology and professional jargon has been created and adopted for use in the recent years. It is not only important to comprehend the meaning of terms, but also understand how they contrast and synergize amongst each another. This paper serves to address the need of building a consistent vocabulary for the newly growing domain of Big Data Analytics. Understanding and adoption of common consistent vocabulary promotes interdisciplinary communication and collaboration and removes entrance barriers for anyone entering the growing world of Big Data Analytics. Using a step-by-step algorithm based on the bibliometric and content analyses of existing peer-reviewed literature, a sample Big Data Analytics vocabulary is built. The approach includes storing terms in the relational database and being able to retrieve and visualize co-related terms thus establishing connections between them. The step-by-step procedure described in the paper involves: 1) collection of data; 2) data manipulation such as elimination of duplicates, identification of synonyms, grammar forms of the same root words and variations in spelling; 3) calculation of frequency of use of the same term; and finally 4) generation of various reports including most frequently used terms per paper or per narrower category, or - in future release – identification of most likely category based on the combination of co-located terms. For this current proof of concept effort, due to complexities and exceptions when dealing with natural language (English), some steps of this process cannot be fully automated, and hence require manual verification or adjustment, although considerable effort was made to minimize the amount of human intervention. The procedure can be repeated periodically with relative ease to observe and report possible changes in the dynamic field of Big Data Analytics and discover newly created vocabulary. Big Data Analytics was chosen for this project because of its characteristics of a not yet thoroughly documented but fast growing field with critical mass of published works already accumulated. This paper hopes to help with creation of educational materials and demarcation of the domain, while encouraging full research coverage in Big Data Analytics, by promoting discovery and articulation of common principles and solutions.**

Keywords—Big data analytics; domain; controlled vocabulary; keyword; content analysis

I. INTRODUCTION

The explosion of number of publications and software products in recent years associated with the term "Big Data Analytics" signals to the scholarly world that the new domain has likely been born. While there exist multiple definitions on what exactly constitutes the domain, it would suffice here to cite Tennis [1] that "to the casual reader … a domain… could be an area of expertise, a body of literature or even a system of people and practices working with a common language". A common language there serves as a vital unifying factor keeping the system whole by tying together people, accumulated knowledge and perspective research areas. Understanding and adoption of common vocabulary promotes interdisciplinary communication and collaboration and removes entrance barriers for newcomers.

Controlled vocabularies are commonly built and maintained by information specialists [2] for specific established subject areas [3] and are visible to the outside world. One such well-respected source is the Library of Congress that maintains Subject Headings (LCSH) and Classification (LCC) schemes [4].

However, finding prebuilt controlled vocabulary for a newly growing domain may not be possible. This paper aims to address the need of building a consistent vocabulary for Big Data Analytics that can fill the void until the official sources catch up.

As my previous research shows [5], the term "Big Data Analytics" itself is relatively new. It appears in scholarly literature only starting from 2010, while a vast amount of new terminology and professional jargon has since been created and adopted for use making it difficult to comprehend the meaning of each term and inter-relate them all.

Still, while the name itself is novel, the necessity to sift through and process massive unstructured data is not. Handfield [6] traces the origin of data analytics to mid-1950. Yet, a recent "proliferation of new software capabilities, accumulated knowledge, jargon and concepts could easily confuse even those who have been involved in analytics at the earlier stages" [5]. It is not only important to comprehend the meaning of each term, but also understand how they contrast and synergize amongst each another.

This paper describes an algorithm and demonstrates a prototype software logic with which Big Data Analytics vocabulary can be built based on the bibliometric and content analyses of existing peer-reviewed literature. It hopes to help with creation of educational materials and demarcation of the domain, while encouraging full research coverage in Big Data Analytics, by promoting discovery and articulation of common principles and solutions. One other potential benefit of outlying the algorithm and thus facilitating repeating its steps periodically with relative ease is that it can let us create

snapshots of the domain's depth and breadth (per Tennis, the intension and extension of the domain [1]) at time intervals, thus tracing the domain's development and having an insight into its future. We then can report possible changes in the dynamic field of Big Data Analytics and discover newly created vocabulary. Also this approach can be extended to apply to other young and fast growing domains that are not yet thoroughly documented but have critical mass of published works already accumulated.

II. Method

Prior to starting data collection it has been confirmed that LCSH and LCC have no existing controlled vocabulary lists searching for the phrase "Big Data Analytics". The quest to produce the missing vocabulary in lieu of authoritative source, then has then been based on Jank's statement [7], "Quite often … authors, scholars, and indexers are more in tune with prominent, current terminology that has not yet made it into thesauri. In these instances, search tags are most commonly located in these fields: *Keyword, Identifier*". Therefore, the plan has been formed to collect a list of keywords from the papers that would demonstrate their dedication to the topic of Big Data Analytics. By reviewing, analyzing and streamlining this list, the goal has been to come up with a draft vocabulary common to those writing on the subject, concentrating on the words and phrases that have been used most often.

Analyzing properties of documents retrieved from scholarly research databases together with other bibliometric techniques has been "an accepted method in the sociology of science" [8]. This method of data collection is both valid and highly reliable as "the data can be collected unobtrusively from readily accessible published records of scholarly communication and thus can be easily replicated by others" [9].

The described method includes storing terms in the relational database and being able to retrieve and visualize co-related items thus establishing connections between them. The step-by-step procedure involves: 1) collection of data; 2) data manipulation such as elimination of duplicates, identification of synonyms, grammar forms of the same root words and variations in spelling; 3) calculation of frequency of use of the same term; and 4) finally, generation of various reports including most frequently used terms or - in future – identification of the combination of co-located terms. Due to complexities and exceptions when dealing with natural language some steps cannot be fully automated and hence require further content analysis to make some verifications or adjustments, although considerable effort has been made to minimize the amount of manual work.

This project has intentionally been limited in scope to serve as a proof of concept to verify and illustrate the selected methodology – no attempt has been made at this point to find all contained terms. As such, the data has been retrieved from a single database only and with a relatively small dataset, the choice of software tools (Microsoft Windows environment utilizing Microsoft Office 2010 utilities, such as VB macros inside Microsoft Excel spreadsheet and Microsoft Access database) has arbitrarily been made based on convenience and the author's familiarity with them, with little or no concerns to performance of the automated steps, clarity of presentation and adherence to the best software engineering practices. All these limitations will be taken into account in the follow up refinement of this approach.

A. Step 1: Data collection

Elsevier Scopus has been selected for conducting the search, as this database 1) is "the largest database for multidisciplinary scientific literature existing on the market… from all areas of knowledge" [10]; and 2) has a convenient post-analysis tool ("a gold mine for scientometrists" [11]) that allows to download the search results in the spreadsheet format (CVS) for subsequent post processing.

The search of Scopus has been performed using the "field-specific searching" [7] technique on a combination of "Article Title, Abstract, Keywords" query – Scopus's top most recommended search field choice. Such field searching comparing to the full-text search allows concentrating on works where "big data analytics" is more likely a principal topic of discussion and weeding out the accidental noise. As such, a Scopus search "TITLE-ABS-KEY ("big data analytics")" has been performed on March 29, 2015 (Fig. 1) and has returned 422 results.

B. Step 2: Elimination of duplicate and empty records

Using Scopus's post-processing tool, the results have been downloaded into Microsoft Excel spreadsheet. The next step has been sorting the results by title and elimination of duplicate records. In this set six duplicates have been found and removed, reducing the total number of results to 416. From this updated set a result number (serving as a unique identifier of a particular written work, automatically generated by Scopus) and a list of keywords associated with each record (two Excel columns), have been copied to a new spreadsheet. Since further manipulations have to be done with keywords, retrieved records with no keywords are eliminated.

C. Step 3: Generation of alphabetic keyword list

In a resulting spreadsheet, keywords for each retrieved work are written in a single cell as a list separated with semicolons (';'). To make this list more manageable for further analysis, the next task has been to restructure this list with each keyword in a cell by itself. This has been achieved using Microsoft Excel's build-in text manipulation tools. Result of running these tools has been a table where each row has represented a unique retrieved record with first column showing its identifier and the remaining variable amount of columns listing the keywords associated with that record (from the previous elimination step it is guaranteed to have at least one keyword per row, i.e., the minimum number of columns in each row is two).

A custom macro shown in Fig. 2 then transposes the keywords vertically, so that all keywords are lined up in a single column with a corresponding result number in the column to the left. That table then can easily be resorted in the order of keywords rather than paper numbers, as has been the sort order of the original spreadsheet.

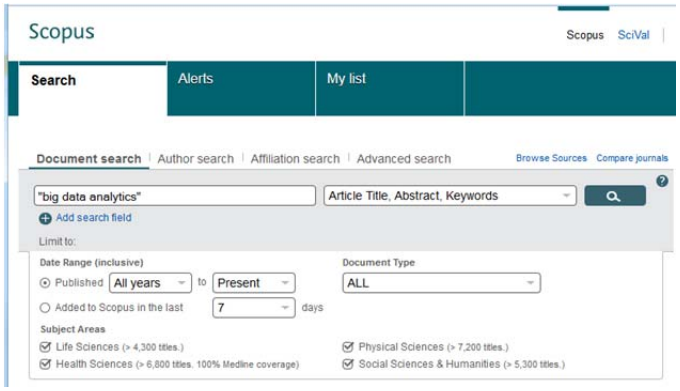

Fig. 1. Step 1 - entering search query in the Scopus database.

D. Step 4: From keywords to vocabulary terms

The goal of the next step has been to eliminate inconsistencies in representation of the keywords from one retrieved result to another. Since this step could involve possible changes to original keywords, rather than modifying them in place, a new column called "Term" has been added to the right of the original "Keyword" column and the keywords has been copied there. This has been done for reproducibility and keeping track of the changes made in case a step back would be needed Meanwhile, Microsoft's Excel built-in TRIM() function cleans up the new column for the occasional extra spaces at the beginning and/or at the end. The results have been resaved as values (not Excel formulas) and resorted by new column order.

The new column then has been manually examined from the top to the bottom to eliminate differences in spelling, using the following set of rules:

a) When the same keyword has been spelled using different combination of uppercase and lowercase letters, replace all such occurrences of uppercase with the lowercase letters.

b) When the same complex term has been spelled with either separate words, combined single word or words connected by a dash, either convert them all to the most common spelling or in some not-so-obvious cases, list in a single cell multiple spellings separated with "/".

c) Fix misspelled words when there are no doubts that they are not correct.

d) When the term is listed both as an abbreviation and spelled-out version, replace a term with a combination of spelled out version following abbreviation in parentheses, i.e., *"Big Data Architecture Framework (BDAF)"*.

e) For the different forms of the same word, combine them all together separated by "/", i.e., *"model/models/modeling"*.

```
Sub Columns_to_Column()

    Dim sht1 As Object
    Dim i As Integer
    Dim j As Integer
    Dim k As Integer

    Set sht1 = Worksheets("Sheet1")

    k = 416

    For j = 3 To 32
        For i = 1 To 422
            sht1.Cells(k + i, 2) = sht1.Cells(i, j).Value
            sht1.Cells(k + i, 1) = sht1.Cells(i, 1).Value
        Next i

        k = k + i - 1
    Next j

    sht1.Range("c1:af416").ClearContents
End Sub
```

Fig. 2. Custom written Microsoft Visual Basic macro to move keywords into a single column with associated paper number to the left.

To make frequency count more accurate, complex multi-word terms have been further examined to determine whether they contain inside them smaller words or phrases that have already been found to be separate terms. If so, these sub-terms have been added to a list with the same referenced record number as in the bigger term. For example, when encountered term "big data analytics", since terms "big data" and "analytics" have already been listed separately, two additional entries for each of them have been created as shown in Table 1. This step has required multiple intermediate resorting in the order of terms.

E. Step 5: Final consolidation and reporting

The output of Step 4 has been imported into Microsoft Access for this final step and saved in a table called "Dictionary" shown in Fig. 3. To check for and eliminate duplicates accidentally introduced in the previous steps, the following SQL query has been run while deleting duplicates until it has returned no results.

TABLE I. A FRAGMENT OF A SPREADSHEET SHOWING CREATION OF NEW OCCURRENCES OF A TERM FROM A BIGGER PHRASE (ADDED ROWS ARE HIGHLIGHTED YELLOW)

OccurredIn	Keyword	Term
…	…	…
219	Big data analytics	big data analytics
219	Big data analytics	big data
219	Big data analytics	analytics
…	…	…

Table: **Dictionary**

Columns

Name	Type	Size	Description
keyword	Text	255	Original keyword spelling from Scopus
term	Text	255	Final term spelling after consolidation operations
occuredIn	Integer	2	Generated paper number as retrieved from Scopus

Table Indexes

Name	Number of Fields	
PrimaryKey	2	
	Clustered:	False
	DistinctCount:	2104
	Foreign:	False

Fig. 3. Content of a Dictionary table created in Microsoft Access.

```
SELECT Dictionary.occurredIn, Dictionary.term,
       Count(*) AS Expr1
FROM Dictionary
GROUP BY occurredIn, term
HAVING Count(*)  > 1;
```

Once the duplicates have been eliminated, Microsoft Access allows indexing the newly created table with combined "term" and "occurredIn" columns as a primary key.

Final Dictionary table has contained 2104 records. From it, a variety of reports have been generated that are being discussed in the next section.

III. Results

As the primary goal of this research has been the generation of Big Data Analytics controlled vocabulary, the first report that has been run from the Access database has produced an alphabetic list of terms (with alternate spellings where present) that has consisted of 197 distinct entries if terms that have appeared only once were ignored. This query is written in SQL as:

```
SELECT term,
       Count(occurredIn) AS Frequency,
       format(Frequency/416,"percent") AS
            ["% of sample records"]
FROM dictionary
GROUP BY term
HAVING Count(occurredIn)>1
ORDER BY term;
```

If no restriction is made (removing HAVING clause from the above SQL statement), the vocabulary list would grow to 904 entries – likely subject to a big noise. A separate discussion, thus, is needed as to what frequency percentage or number shall be considered as a cutoff for inclusion into a domain vocabulary list.

To help determine such cutoff and visualize frequency distribution, another query has been run to calculate and display frequent terms (appearing in document keywords more than once). The number 416 has been used in percentage calculations as the number of unique papers retrieved by Scopus. That would mean that a hypothetical term appearing in all 416 retrieved results simultaneously, would have a 100% frequency count – the maximum possible. In SQL statement, this query is:

```
SELECT term,
       COUNT(occuredIn) AS Frequency,
       FORMAT( Frequency / 416, "Percent") AS
            ["% of sample records"]
FROM dictionary
GROUP BY term
HAVING Count(occurredIn) > 1
ORDER BY Frequency DESC, term;
```

Table 2 shows results of this query in its default descending order from most to least common for the terms with the frequency count exceeding five (top 45). Complete results are available from the author upon request.

IV. Discussion

These results have a somewhat subjective frequency counts as while some terms have been manually grouped together, the others, appearing slightly more distinct, have not (for example, "*model*" and "*modeling*" have been grouped while "*mobile*" and "*mobility*" have not). It has often been a thin line to cross, and the approach has been chosen to "undergroup" rather than "overgroup". In some cases, the meaning of the abbreviation has not been provided or recognized and that has prevented grouping of possibly related terms.

Scopus CSV download feature and Microsoft Excel tools and macro language have been a big help in automating some portions of this research. However, while unexpectedly spending a significant amount of time on manually performing Step 4, I believe now that at least some portions of it could next time be automated as well with the use of another custom macro, especially if dealing with bigger set of data.

At the end of the day, the proposed methodology worked, as it produced a viable and potentially useful vocabulary for a fast growing domain of Big Data Analytics. While validity of this work could improve if data are collected from the multiple sources (considering next time to query other databases as well, such as IEEE Xplore and ACM Digital Library), its reliability, on the other hand, is secured by reproducibility of the documented steps. All-in-all, chosen methodology has shown to be a successful proof of concept. The follow-up steps would include adding data sources; making the procedure more robust; increasing automation and reducing manual labor as much as possible, thus decreasing the possibility of errors; providing additional reports that would allow new insights into the Big Data Analytics domain, as for example, identification of subgroups within domain based on the combination of frequent co-related terms.

TABLE II. DATA DICTIONARY IN BIG DATA ANALYTICS SORTED BY FREQUENCY OF USE

Dictionary Query		
term	***Frequency***	***Percentage in sample records***
big data	184	44.23%
analytics	118	28.37%
data analytics	84	20.19%
big data analytics	75	18.03%
cloud	47	11.30%
MapReduce	35	8.41%
analysis	33	7.93%
cloud computing	31	7.45%
management	31	7.45%
Hadoop	27	6.49%
Information	24	5.77%
service/services	19	4.57%
security	17	4.09%
Database/Databases	15	3.61%
intelligence	14	3.37%
model/models/modeling	14	3.37%
performance	13	3.13%
business	12	2.88%
data mining	12	2.88%
architecture/architectures	10	2.40%
Information technology/technologies	10	2.40%
machine learning	9	2.16%
algorithm/agorithms	8	1.92%
Business intelligence (BI)	8	1.92%
data analysis	8	1.92%
distributed system/systems	8	1.92%
framework	8	1.92%
Healthcare	8	1.92%
internet of things (IoT)	8	1.92%
research	8	1.92%
visualization	8	1.92%
Benchmark/benchmarking	7	1.68%
cluster/clusters/clustering	7	1.68%
Data warehouse/warehousing	7	1.68%
Database management	6	1.44%
Database management system (DBMS)	6	1.44%
enterprise	6	1.44%
file system/systems	6	1.44%
On-line analysis/Analytical Processing (OLAP)	6	1.44%
Ontology/Ontologies	6	1.44%
parallel computation/computing	6	1.44%
platform/platforms	6	1.44%
Predictive analytics	6	1.44%
privacy	6	1.44%
SQL	6	1.44%

V. CONCLUSION

While significant portion of this paper has been devoted to justifying and proving selected methodology, the procedure described in here has not been the goal in itself. Rather, it is a device aiming to increase our knowledge of the Big Data Analytics domain. The produced results are both semi-expected and thought-provoking. For example, we predictably have been shown that the various combinations of terms "big", "data" and "analytics" are at the top of the frequency list. An interesting phenomenon is that the notions of cloud and cloud computing have been at the top of the list signaling that there is a significant intersection of research interests between these two domains. Among other most frequent terms are the specific software tools Hadoop and MapReduce that indeed are most commonly being identified with Big Data Analytics. Effective management, security and performance appear to be the most popular research problems. The other common terms help identify major functions, methodology and concepts of the Big Data Analytics domain (services, architectures, databases, data mining, machine learning, etc.).

As Wildemuth [12] rightly warned, "in essentially every study, data obtained from documents or artifacts … need to be analyzed in combination with data obtained using other methods". This paper provides one of the snapshots under specific angle into Big Data Analytics - a building block among many on which our full understanding of the domain will eventually stand. It also creates a core vocabulary useful for all the purposes set at start. Yet, while the domain itself keeps evolving, so is the need to keep pace with its development – and thus, the best ending that comes to mind is "To Be Continued…"

ACKNOWLEDGMENT

Special thanks to David Jank, Ph.D. from Palmer School, Long Island University for review and encouragements.

REFERENCES

[1] J. T. Tennis, Two Axes of Domains for Domain Analysis, Knowledge Organization. 30(3/4), pp. 191-195, 2003.

[2] M. J. Bates, The invisible substrate of information science. Journal of the American Society for Information Science 50, #12, pp. 1043-1050, 1999.

[3] H Chu, Information representation and retrieval in the digital age, 2nd ed., Information Today, Medford, NJ, p. 54, 2010.

[4] Cataloging And Acquisitions (Library Of Congress). Loc.gov. September 1, 2016. Available: https://www.loc.gov/aba/ .

[5] L. Turiy, Classification of research efforts in dynamic/big data analytics. 12th International Conference & Expo on Emerging Technologies for a Smarter World (CEWIT), October 19-20, 2015; Melville, NY. IEEE. doi: 10.1109/CEWIT.2015.7338171.

[6] R. Handfield, A Brief History of Big Data Analytics, September 26, 2013. Available: http://iianalytics.com/research/a-brief-history-of-big-data-analytics .

[7] D. Jank, Database Gymnastics. How to get out what the publishers put in …. and how to know what they put in! Unpublished. Palmer School of Library and Information Science, LIU – C.W. Post Campus, NY, 2010.

[8] C. L. Borgman and J. Furner, Scholarly communication and bibliometrics, Edited B. Cronin, Annual Review of Information Science and Technology, Information Today, Medford, NJ, Vol. 36, pp. 3-72, 2002.

[9] D. Zhao, A comparative citation analysis study of Web-based and print journal-based scholarly communication in the XML research. Dissertation, Florida State University, 2003.

[10] F. De Moya-Anegón, Z Chinchilla-Rodríguez, B. Vargas-Quesada, E. Corera-Álvarez, F. J. Muñoz-Fernández, A. González-Molina and V.

Herrero-Solana, Coverage analysis of scopus: A journal metric approach. Scientometrics, 73(1), pp. 53-78, 2007.

[11] P. Jasco, Scopus, Péter's Digital Reference Shelf. September 2004. Available: http://www.galegroup.com/servlet/HTMLFileServlet?imprint=9999&region=7&fileName=reference/archive/200409/scopus.html.

[12] B. M. Wildemuth, Applications of Social Research Methods to Questions in Information and Library Science, Libraries Unlimited, Santa Barbara, CA , p. 161, 2009.