International Journal of Research in Advanced Engineering and Technology
ISSN: 2455-0876; Impact Factor: RJIF 5.44
www.engineeringresearchjournal.com
Volume 2; Issue 4; July 2016; Page No. 11-14

Big data analytics in engineering and science curriculum restructuring using singular value decomposition (SVD)

Ozor Godwin O, Nwobodo Lois O, Oleka Chioma V

Computer Engineering, Enugu State University of Science and Technology, Enugu, Nigeria

Abstract
With the availability of huge data in the internet and other web supported devices. It was reported by other researchers that lesser percentage of the data were being used. Engineering and Science curriculum need extensive review and restructuring to improve employability of the graduates and also, to meet the needs of the industry. In this paper, critical areas like: communication skills; problem solving skills; entrepreneur skills; environmental awareness; lifelong learning; information management; teamwork; ethics and moral were considered for course(s) to be taught in the faculty of engineering and science. Singular value decomposition was applied to align the component of effective real-world training pattern to bridge the gap between academic and industry. Matlab was used to simulate the model.

Keywords: Big data, SVD, curriculum, engineering and science

1. Introduction

The use of internet is now a commonplace in our daily lives. The advent of the internet, followed by the www boom increased our mailing systems, education systems, banking systems, retailing and entertainment, leading to storage and transmission of voluminous data. As we enter into the information age, data are being generated by variety of sources other than servers and human, such as sensors embedded into phones and wearable devices, surveillance cameras and scanners. Taking into consideration the sources of the data both in physical universe and in digital universe, the physical and digital universe was created by everyone using a digital camera, by the more than 2 billion people and millions of enterprises living their lives and doing their work online, and by the millions of sensors and communicating devices sending and receiving data over the Internet. But unlike the physical universe, the digital universe is created and defined by software, a man-made construct. It is defined by software that analyzes this ever-expanding universe of digital data, finding the hidden value and new opportunities to transform and enhance the physical world – keeping the academic research and laboratories experiments. And it is software that will both create new opportunities and new challenges for us as we try to extract value from the digital universe that we have created. Today, the digital universe has reached a number of new thresholds: The data coming from embedded systems (e.g., MP3 players, traffic lights, MRI scanners) has grown to a level where it's starting to challenge established practices in datacenters; the migration to digital entertainment – movies and TV – is almost complete; and metadata (e.g., the data added to your email message describing when it was created), once tightly coupled with the data it describes, has grown into a category in and of itself, the fastest-growing subcategory of the digital universe [1].

Activities of science and engineering, through embedded systems and multimedia encouraged the growth of data and it was predicted to cater for 10 percent of the accumulated data by 2020 [3]. Hence the genealogy of data growth and transformation was categorized into four tiers. The first was when digital camera technology replaced film; the second, when analog telephony went digital; and the third, when TV went digital. Now comes a fourth growth spurt the migration of analog functions monitoring and managing the physical world to digital functions involving communications and software telemetry. Call it the advent of the Internet of Things (IoT) [2]. Fed by sensors soon to number in the trillions, working with intelligent systems in the billions, and involving millions of applications, the Internet of Things will drive new consumer and business behavior that will demand increasingly intelligent industry solutions.

The aggregation of engineering and science data in the internet was enough to streamline engineering and science contents relevant to industrial revolution. Many experiment and simulation were done in the past, whether published or unpublished, but the power of content analysis with respect to recent works are yet to be explored. The Big data analytics through singular value decomposition was considered in this paper as the better option to build a viable and lifelong engineering and science content for undergraduates.

2. Material and Methodology

Matrix factorization was used to develop the easy method of determining relevant information about the item for the model. The items are tagged 'terms' while the webpages are tagged 'documents'. The terms for this research are: communication skills, (CS); problem-solving skills, (PS); entrepreneur skills, (ES); environmental awareness, (EA); lifelong learning, (LL); information management, (IM); teamwork, (Tw); ethics and moral(EM). Four documents and terms are used for the research modelling.

Table 1: Frequency of each item in the sampled four documents (DcX)

Document/Term	Dc1	Dc2	Dc3	Dc4
CS	1	1	0	3
PS	3	2	5	1
ES	0	2	1	0
EA	0	0	1	2
LL	2	0	0	1
IM	1	1	2	0
Tw	0	2	4	1
EM	2	0	1	5

The associated term by document matrix $G \in R^{m\times n}$ is constructed directly from the table 1.

$$G = \begin{pmatrix} 1\,1 & 0\,3 \\ 3\,2 & 5\,1 \\ 0\,2\,1\,0 \\ 0\,0\,1\,2 \\ 2\,0\,0\,1 \\ 1\,1\,2\,0 \\ 0\,2\,4\,1 \\ 2\,0\,1\,5 \end{pmatrix}$$

The entries in G, denoted g_{ij}, represent the frequency in which the i^{th}term occurs in the j^{th} document. The matrix G, is inherently sparse because it would be unusual to find every term in every document. This is very powerful observation, since computational algorithms and computer technology that exploits this sparsity can be used to speed up the search process. If a query vector is assembled in the form $\boldsymbol{p} = (p1, p2, p3, \dots, pm)^G$, where the entries $pi = 1$ if term G_i, $i = 1, \dots, m$ appears in the query, otherwise $pi = 0$. A weighted might also be applied to these entries, p_i, for simplicity emphasis was to measure how well the query matches $Dcj, j = 1, \dots, n$. we check how close the query vector p is to the j^{th}column of the term-by-document matrix.
We measure how close the p is to $G *_j$ by computing

$$\cos \alpha j = \frac{P^G G * j}{|P||G * j|} \quad j = 1, \dots, n.$$

The resulting vector $(\cos \alpha 1, \cos \alpha 2, \dots, \cos \alpha n)$ provides information that the search engine can use to rank the relevance of each document of each document relative to the engineering and science contents query.

```
Terms = {'CS','PS','ES','EA','LL','IM','Tw','EM'}
G = [1 1 0 3; 3 2 5 1;0 2 1 0;0 0 1 2;2 0 0 1;1 1 2 0;0 2 4 1;2 0 1
5];
display(G)
% vectors corresponding to the search terms
p = [0; 0; 1; 0; 1; 1; 0; 0];
n = size (G,1);
m = size(G,2);
fprintf ('Terms Selected:  ');
hand =  waitbar(1,'Please Wait....');
for i = 1:n
   if (p(i,1) == 1)
      fprintf('%s ',Terms{i});
   end
end
fprintf('\n')
fprintf('\n')
fprintf('\n')
for j = 1:m
   fprintf('cos alphaj = %g\n',p'*G(:,j)/norm(p)/norm(G(:,j)));
end
```

The output of the query is as shown below

```
Terms = 'CS' 'PS' 'ES' 'EA' 'LL' 'IM' 'Tw' 'EM'
G =     1 1 0 3
        3 2 5 1
        0 2 1 0
        0 0 1 2
        2 0 0 1
        1 1 2 0
        0 2 4 1
        2 0 1 5
Terms Selected: ES LL IM
cos alphaj = 0.39736
cos alphaj = 0.46291
cos alphaj = 0.25
cos alphaj = 0.090167
```

The outpout shows that the second document is most relevant to our search. It does not contain the term Lifelong (LL) but it contains (ES) twice and (IM) once. The research process seem straightforward, but we quickly point out that the differences, ambiguity and variations on vocabulary, as well as subtleties in the indexing processes, potentially lead to substantial amount of 'noise' evident in the term- by-document matrix G. Hence the SVD provides us with an ideal way of filtering out noise by using a low rank truncation.
Singular Value Decomposition model expression is: $A = U\Sigma V$ for the case $m > n$. in general matrix $A \in R^{m\times m}$ refer to the fundamental theorem states that orthogonal matrices $U \in R^{m\times m}$ and $V \in R^{n\times n}$
In applications the SVD to document retrieval, the columnd U_j of U are commonly referred to as term vectors and the columns V_j of V are referred to as document vectors. Latent semantic indexing (LSI) uses the truncated low rank approximation G_l in place of **G** as a way to filter the noise associated with the ambiguity in the vocabulary.

3. Results

The results of the Singular Value Decomposition of complex and huge data were simulated and presented as follows:

```
>> [evecs, evals] = eig(G'*G, 'vector');
>> display(evals)

evals =

    2.9828
    6.7472
   31.0221
   81.2478

>> [evals_sorted, perm] = sort(evals,'descend')

evals_sorted =

   81.2478
   31.0221
    6.7472
    2.9828


perm =

     4
     3
     2
     1
```

```
>> G = [1 1 0 3; 3 2 5 1;0 2 1 0;0 0 1 2;2 0 0 1;1 1 2 0;0 2 4 1;2 0 1 5];
>> display(G);

G =

     1     1     0     3
     3     2     5     1
     0     2     1     0
     0     0     1     2
     2     0     0     1
     1     1     2     0
     0     2     4     1
     2     0     1     5

>> rank(A)

ans =

     2
```

```
>> evecs_sorted =evecs(:, perm)

evecs_sorted =

    0.3995   -0.1472    0.8807    0.2077
    0.3260    0.2904   -0.2994    0.8484
    0.6891    0.5278   -0.1101   -0.4842
    0.5093   -0.7845   -0.3502   -0.0508

>> H1 = evecs_sorted

H1 =

    0.3995   -0.1472    0.8807    0.2077
    0.3260    0.2904   -0.2994    0.8484
    0.6891    0.5278   -0.1101   -0.4842
    0.5093   -0.7845   -0.3502   -0.0508

>> H1'*H1

ans =

    1.0000   -0.0000    0.0000    0.0000
   -0.0000    1.0000   -0.0000    0.0000
    0.0000   -0.0000    1.0000   -0.0000
    0.0000    0.0000   -0.0000    1.0000
```

```
>> S = diag(sqrt(evals_sorted))

S =

    9.0138         0         0         0
         0    5.5697         0         0
         0         0    2.5975         0
         0         0         0    1.7271

>> U1 = G*H1*inv(S)

U1 =

    0.2500   -0.3968   -0.1807    0.5233
    0.6440    0.3579    0.4398   -0.0880
    0.1488    0.1990   -0.2729    0.7021
    0.1894   -0.1869   -0.3120   -0.3392
    0.1451   -0.1937    0.5433    0.2112
    0.2334    0.2152    0.1390    0.0507
    0.4346    0.3425   -0.5349   -0.1685
    0.4476   -0.6623   -0.0383   -0.1868

>> U2 = null(G')
```

```
U2 =

   -0.3812   -0.0628    0.1111   -0.5585
   -0.2206   -0.3296   -0.3074   -0.0644
    0.3051    0.0017   -0.3511    0.3930
    0.4511    0.1402   -0.5533   -0.4330
    0.6713   -0.0047    0.3528   -0.1627
   -0.1447    0.9233    0.0463   -0.0417
    0.1784   -0.1023    0.5798   -0.0285
   -0.0776    0.0689    0.0296    0.5594

>> U = [U1 U2]

U =

    0.2500   -0.3968   -0.1807    0.5233   -0.3812   -0.0628    0.1111   -0.5585
    0.6440    0.3579    0.4398   -0.0880   -0.2206   -0.3296   -0.3074   -0.0644
    0.1488    0.1990   -0.2729    0.7021    0.3051    0.0017   -0.3511    0.3930
    0.1894   -0.1869   -0.3120   -0.3392    0.4511    0.1402   -0.5533   -0.4330
    0.1451   -0.1937    0.5433    0.2112    0.6713   -0.0047    0.3528   -0.1627
    0.2334    0.2152    0.1390    0.0507   -0.1447    0.9233    0.0463   -0.0417
    0.4346    0.3425   -0.5349   -0.1685    0.1784   -0.1023    0.5798   -0.0285
    0.4476   -0.6623   -0.0383   -0.1868   -0.0776    0.0689    0.0296    0.5594
```

```
>> U'*U

ans =

    1.0000   -0.0000    0.0000    0.0000         0   -0.0000    0.0000         0
   -0.0000    1.0000    0.0000    0.0000    0.0000   -0.0000   -0.0000   -0.0000
    0.0000    0.0000    1.0000   -0.0000    0.0000   -0.0000   -0.0000    0.0000
    0.0000    0.0000   -0.0000    1.0000   -0.0000   -0.0000   -0.0000    0.0000
         0    0.0000    0.0000   -0.0000    1.0000   -0.0000    0.0000    0.0000
   -0.0000   -0.0000   -0.0000   -0.0000   -0.0000    1.0000    0.0000   -0.0000
    0.0000   -0.0000   -0.0000   -0.0000    0.0000    0.0000    1.0000   -0.0000
         0   -0.0000    0.0000    0.0000    0.0000   -0.0000   -0.0000    1.0000
```

```
>> Sigma = [S; zeros(4,4)]

Sigma =

    9.0138         0         0         0
         0    5.5697         0         0
         0         0    2.5975         0
         0         0         0    1.7271
         0         0         0         0
         0         0         0         0
         0         0         0         0
         0         0         0         0

>> U*Sigma*H1'

ans =

    1.0000    1.0000    0.0000    3.0000
    3.0000    2.0000    5.0000    1.0000
   -0.0000    2.0000    1.0000    0.0000
    0.0000    0.0000    1.0000    2.0000
    2.0000   -0.0000    0.0000    1.0000
    1.0000    1.0000    2.0000    0.0000
    0.0000    2.0000    4.0000    1.0000
    2.0000    0.0000    1.0000    5.0000
```

```
>> G-U*Sigma*H1'

ans =

   1.0e-14 *

   -0.0222         0   -0.0333         0
   -0.0444         0    0.0888   -0.1554
    0.0167         0   -0.0222   -0.0271
   -0.0625   -0.0333         0         0
   -0.0444    0.0278   -0.0527   -0.0444
    0.0111   -0.0222    0.0444   -0.0413
   -0.0777   -0.0888    0.0444   -0.0888
   -0.1332   -0.0167   -0.0444         0
```

The results make use of the fundamental theorem of orthogonal matrices of A-UΣH1' to show the randomness and availability of required data in the space.

4. Conclusion

Big data analytics is a better tools for viability, completeness of curriculum restructuring especially in engineering and science. It explained that much data are in the space, there are little or no need for new experiment without harnessing the huge data of past researched work. With proper analytical tools, sound engineering and science education will be realized.

5. References

1. V. Turner JF, Gantz D. Reinsel Minton S. The Digital Universe of Opportunities: Rich Data and the Increasing Value of the Internet of Things. International Data Corporation, White paper, April 2014
2. Atzori L, Iera A, Morabito G. The Internet of Things: A survey Computer Networks 2010; 54:2787-2805.
3. pixuffle.net, The Importance of Computers in Our Dailly Lives, [ref¬erence made 20.05.2015]. Available at: http://www.pixuffle.net/the-importance-of-computers-in-our-daily-lives/
4. Thomas Davenport, Jill Dyché. Internal Institute for Analytics, Big Data in Big Companies, [reference made 13.04.2015]. Available at: http://www.sas.com/content/dam/SAS/en_us/doc/whitepaper2/big¬data-bigcompanies-106461.pdf
5. Juniper networks. Introduction to Big Data: Infrastructure and Networking Considerations Leveraging Hadoop-Based Big Data Archi¬tectures for a Scalable, High-Performance Analytics Platform, [refer¬ence made 13.04.2015]. Available at: http://www.juniper.net/us/en/local/pdf/whitepapers/2000488-en.pdf
6. Kevin Normandeau, Inside Big Data, Beyound Volume, Variety and Ve¬locity is the Issue of Big Data Veracity, [reference made 07.04.2015]. Available at: http://insidebigdata.com/2013/09/12/beyond-volume-variety-velocity-issue-big-data-veracity/
7. Jared Dean, Big Data, Data Mining and Machine Learning, Value Creation for Business Leaders and Practitioners, Wiley, 2014, 3-5.